UC Merced

UC Merced Electronic Theses and Dissertations

Title

Identifying Heterogeneity in Protein Crystal Structures for IRF3 Drug Discovery and Enzyme Design

Permalink

https://escholarship.org/uc/item/0c43d4rr

ISBN

9798173793768

Author

Rodriguez, Joshua Michael

Publication Date

2026-06-16

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, MERCED

Identifying Heterogeneity in Protein Crystal Structures for IRF3 Drug Discovery and Enzyme Design

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Chemistry and Biochemistry

by

Joshua M. Rodriguez

Committee in charge:

Professor Patricia LiWang, Chair
Professor Shahar Sukenik
Professor Maria-Elena Zoghbi
Professor Michael Thompson

2026

Copyright

Joshua M. Rodriguez, 2026

All rights reserved.

The dissertation of Joshua M. Rodriguez is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

(Professor Shahar Sukenik)

(Professor Maria-Elena Zoghbi)

(Professor Michael Thompson)

(Professor Patricia LiWang, Chair)

University of California, Merced

2026

DEDICATION

To my parents, whose unwavering support and guidance allowed me to explore my passion in science. And to my brothers, whose steadfast camaraderie has made every triumph sweeter and every struggle more bearable.

TABLE OF CONTENTS

Signature Page . . . iii

Dedication . . . iv

Table of Contents . . . v

List of Figures . . . vii

List of Tables . . . viii

Acknowledgements . . . ix

Vita and Publications . . . xi

Abstract . . . xii

Chapter 1 Introduction . . . 1
1.1 IRF3 and innate antiviral immunity . . . 2
1.2 The IRF3 activation cycle . . . 3
1.3 The Structure of IRF3CTD . . . 5
1.4 IRF3 in human disease . . . 8
1.5 Viral antagonism of IRF3 . . . 9
1.6 Chemical tractability of IRF3 and crystallographic fragment screening . . . 10
1.7 Conformational dynamics and the redesign of enzyme function . . . 12
1.8 Dissertation overview . . . 15

Chapter 2 Crystallographic fragment screen of IRF3CTD . . . 17
2.1 Introduction . . . 17
2.2 Results . . . 19
2.2.1 Site 1 includes 13 hits in a buried hydrophobic pocket . . . 19
2.2.2 Three fragments bind site 2, a surface pocket adjacent to the C-terminal tail . . . 20
2.2.3 One fragment binds at a key dimer and adaptor interface . . . 22
2.3 Discussion . . . 23
2.4 Fragment Development . . . 27
2.5 Conclusion . . . 28
2.6 Methods . . . 29

Chapter 3 Biophysical characterization of the proposed RSV NS1–IRF3 interaction . . . 32
3.1 Introduction . . . 32
3.2 Results . . . 34
3.2.1 Analytical Size-Exclusion Chromatography . . . 34
3.2.2 Fluorescence Polarization . . . 34
3.2.3 Biolayer Interferometry . . . 36
3.3 Discussion . . . 38
3.4 Methods . . . 40

Chapter 4 Crystallography-informed computational protein design . . . 44
4.1 Part A: Computational remodeling of an enzyme conformational landscape for altered substrate selectivity . . . 44
4.1.1 Introduction . . . 45
4.1.2 Results . . . 46
4.1.3 Discussion . . . 64
4.1.4 Methods . . . 65
4.2 Part B: Design of efficient artificial enzymes using crystallographically enhanced conformational sampling . . . 74
4.2.1 Introduction . . . 75
4.2.2 Results . . . 77
4.2.3 Discussion . . . 91
4.2.4 Methods . . . 94

Chapter 5 Conclusion . . . 102
5.1 Crystallographic Fragment Screening Opens Doors for Ligand and Drug Development . . . 103
5.2 Computational Design Benefits from Crystallography . . . 105
5.3 Future Applications . . . 106

LIST OF FIGURES

Figure 1.1: Schematic diagram of IRF3 . . . 3
Figure 1.2: Crystal structures reveal IRF3CTD's autoinhibitory mechanism 7

Figure 2.1: Crystallographic fragment screen of IRF3CTD yields 25 hits across eight binding sites . . . 20
Figure 2.2: Site 1 . . . 21
Figure 2.3: Three fragments bind site 2, adjacent to the C-terminal tail . . 22
Figure 2.4: Site 7 . . . 24
Figure 2.5: Sites 1 and 2 contain fragments in close proximity to key interfaces . . . 26
Figure 2.6: Potential routes for fragment development . . . 28

Figure 3.1: NS1 disrupts the pIRF3CTD dimer in size-exclusion chromatography . . . 35
Figure 3.2: Fluorescence polarization shows that NS1 binds IRF3CTD with similar affinity to lysozyme . . . 36
Figure 3.3: BLI with biotin-IRF3CTD and NS1 reveals nonspecific binding of NS1 to the probe surface . . . 37
Figure 3.4: No binding is detected between biotin-NS1 and pIRF3FL . . . 38

Figure 4.1: AAT conformational landscape . . . 48
Figure 4.2: Computational remodeling of AAT conformational landscape by multistate design . . . 50
Figure 4.3: Crystal structures . . . 59
Figure 4.4: Conformational landscape analysis by NMR . . . 61
Figure 4.5: Graphical abstract from the original publication. . . . 75
Figure 4.6: **Scheme 1.** Reaction catalyzed by HG-series Kemp eliminases. 78
Figure 4.7: Ensemble-based computational enzyme design . . . 79
Figure 4.8: Ensemble-based design vs directed evolution . . . 84
Figure 4.9: Crystal structures of HG series Kemp eliminases . . . 85
Figure 4.10: Computational models vs crystal structures . . . 87
Figure 4.11: Ensemble-based design leads to a shift in the conformational ensemble . . . 90

LIST OF TABLES

Table 4.1: Apparent kinetic parameters of *E. coli* AAT and its mutants . . 52
Table 4.2: Conformational equilibrium of AAT variants 55
Table 4.3: Kinetic parameters of Kemp eliminases 82

ACKNOWLEDGEMENTS

I am deeply thankful to my advisor, Dr. Michael Thompson, for both supporting and challenging me throughout my doctoral program. His mentorship consistently led me through the inevitable turbulence of graduate school and has fundamentally shaped how I think as a scientist. I'm grateful for the trust he placed in me to chart my own course, and for knowing when to let me follow a hunch and when to pull me back. I'm also grateful for the lab culture he cultivated, where people from all backgrounds feel welcome and valued. I can only hope to bring the same balance of rigor, generosity, and openness to those I mentor in the future.

I would also like to thank the members of my dissertation committee, whose guidance shaped this work in meaningful ways. I came away from every meeting with a sharpened understanding of my scientific problems and new ideas to explore. I am especially grateful to Dr. Shahar Sukenik for teaching me a remarkably useful set of computational and analytical skills; my work is more rigorous for the ways he challenged me over the years. Dr. Patricia LiWang also consistently pushed me to think beyond the immediate problem and consider the broader significance of my work, and her help in troubleshooting challenging experiments was invaluable. I would also like to acknowledge Dr. Juris Grasis, who served thoughtfully and supportively on my committee for years before unforeseen circumstances required him to step away shortly before my defense. Finally, I want to thank Dr. Maria-Elena Zoghbi for her flexibility and generosity in joining my committee on short notice. I am grateful that she was willing to see me through the final stretch.

I am immensely grateful to Dr. Galen Correy and the Fraser lab at UC San Francisco, whose contributions were instrumental to the success of my fragment screening project. Dr. Correy generously shared his time and lab resources to help me troubleshoot crystallization, and taught me how to carry out the fragment screen from crystal soaking through fragment analysis. I learned an enormous amount from him, and I am fortunate to be continuing that work as a postdoc in the Fraser lab.

I would also like to thank Dr. Roberto Chica at the University of Ottawa, who

provided me with the opportunity to collaborate with his group and gain valuable crystallographic experience early in my graduate career. The two projects that I contributed to shaped my crystallographic skillset and laid the groundwork for my later success on the fragment screening project. Working with him also gave me a deeper appreciation for protein design and biophysics, broadening my scientific perspective in ways that have served me well throughout graduate school.

I have also been fortunate to share a lab with an extraordinary group of people. I am grateful to Luisa Garcia Michel, for being a constant sounding board for my work and for the warmth of her friendship. To Alexander Deary, for all of the brainstorming and speculating, the pushback, and the countless conversations that helped me work through scientific ideas. To Behnoush Seifinoferest, for her willingness to help and her constant encouragement. To Andree Souder, for her thoughtfulness, and for the conversations and laughter that always lift my mood. To Azeem Horani, whose feedback consistently sharpened my work. To Alec Martinez, for holding the lab together through all the work that goes unseen. And to Dr. Alex Wolff, whose generous mentorship has helped shape the way I think about structural biology. Finally, to the undergraduates who worked alongside me on my projects, Vincent Hernandez, Octavio Barron, and Marissa Sweeney: thank you for your dedication and hard work.

VITA

2019 B.S. in Biochemistry, California State University, Fresno

2026 Ph.D., University of California, Merced

PUBLICATIONS

Rakotoharisoa, R. V.; Seifinoferest, B.; Zarifi, N.; Miller, J. D. M.; Rodriguez, J. M.; Thompson, M. C.; Chica, R. A. "Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling". *J. Am. Chem. Soc.* **2024**, 146, 10001–10013.

St-Jacques, A. D.; Rodriguez, J. M.; Eason, M. G.; Foster, S. M.; Khan, S. T.; Damry, A. M.; Goto, N. K.; Thompson, M. C.; Chica, R. A. "Computational Remodeling of an Enzyme Conformational Landscape for Altered Substrate Selectivity". *Nat. Commun.* **2023**, 14, 6058.

ABSTRACT OF THE DISSERTATION

Identifying Heterogeneity in Protein Crystal Structures for IRF3 Drug Discovery and Enzyme Design

by

Joshua M. Rodriguez

Doctor of Philosophy in Chemistry and Biochemistry

University of California Merced, 2026

Professor Patricia LiWang, Chair

The transcription factor IRF3 is central to the innate antiviral immune response and is a therapeutically attractive target in inflammation, antiviral therapy, and cancer. Yet, no experimentally validated small-molecule ligands have been reported for IRF3. Additionally, while many viral antagonists have been reported to interact with IRF3 and disrupt its function, most lack biophysical confirmation of direct binding. Here, we apply crystallographic fragment screening and complementary biophysical methods to address both gaps. First, we present a crystallographic fragment screen of the IRF3 regulatory domain. Pan-dataset analysis across hundreds of ligand-soaked crystals leverages the statistical power of an averaged data set in identifying outliers in single data sets that represent ligands bound to IRF3. We identify 25 fragments bound across eight sites on the IRF3 regulatory domain. We then test a proposed interaction between IRF3 and the respiratory syncytial virus protein NS1 using biophysical assays. Specific binding was not detected, underscoring the need for biophysical validation of IRF3-viral antagonist interactions. We then extend our methodological approach, treating crystallographic data as a statistical ensemble rather than a route to a single structure, to computational protein design. In a pair of collaborative enzyme-design projects, crystal structures that capture conformational heterogeneity served as inputs to

multistate and ensemble-based design strategies. The resulting designed proteins showed tuned conformational equilibria and improved catalytic efficiencies. Together, these projects highlight the value of recovering structural information about heterogeneity that conventional crystallographic methods miss.

Chapter 1

Introduction

X-ray crystallography has long been the dominant method for determining macromolecular structures at atomic resolution. The standard practice produces a single, averaged structure from a cryo-cooled crystal. This provides a static snapshot of the protein, obscuring at least two types of information central to function. The first is conformational heterogeneity, the alternative states a protein samples that often determine catalysis, allostery, and ligand binding. The second is compositional heterogeneity, the partial occupancy of bound ligands or alternative chemical states across the molecules in the crystal. A growing set of crystallographic methods recovers this information from diffraction data. These methods expand what crystallography can address beyond static structure determination.

This dissertation applies these methods to two systems. The first is the antiviral transcription factor IRF3, a therapeutically significant target for which no experimentally validated small-molecule ligands have been reported. Chapter 2 presents a crystallographic fragment screen of the IRF3 regulatory domain, using pan-dataset analysis across hundreds of ligand-soaked crystals to identify low-occupancy ligands that conventional difference-map approaches miss. Chapter 3 complements this with solution-state biophysical analysis, testing whether a proposed viral antagonist of IRF3 binds the protein directly. The second system, addressed in Chapter 4, comprises a pair of collaborative protein-design projects with the Chica group at the University of Ottawa. In these projects, crystal structures that capture conformational heterogeneity are leveraged for protein design

methods that optimize for these different states. The remainder of this chapter establishes the biological, structural, and therapeutic context for these projects, beginning with IRF3 and the innate antiviral immune response.

1.1 IRF3 and innate antiviral immunity

The innate immune system is a critical, early-stage line of host defense against viral infection. Within minutes to hours of pathogen detection, infected cells mount a response that establishes an antiviral state, recruits adaptive immunity, and limits viral replication. At the heart of this response in vertebrates is the interferon (IFN) system, and at the heart of type I IFN induction is interferon regulatory factor 3 (IRF3). IRF3 is a transcription factor that integrates signals from multiple cytosolic and membrane-bound pattern recognition receptors and converts them into a coordinated transcriptional program [1, 2, 3].

IRF3 is constitutively expressed in almost all cell types and is held in a latent, inactive state in the cytoplasm in the absence of infection [1, 4]. This pre-formed pool of IRF3 enables a rapid antiviral response. Upon detection of pathogen-associated molecular patterns (PAMPs) by upstream sensors, IRF3 is activated by phosphorylation, homodimerizes, translocates to the nucleus, and drives transcription of type I interferons along with a smaller set of directly IRF3-responsive genes including the chemokines CCL5 and CXCL10 [5, 2]. Secreted interferons then engage cell-surface receptors in autocrine and paracrine fashion, triggering JAK-STAT signaling that drives the broader transcriptional program of interferon-stimulated genes [6]. The antiviral response that follows is then amplified by a second wave of IRF7-dependent transcription, but the initial transcriptional commitment to interferon induction is made by IRF3 [3]. The functional consequences of this commitment extend beyond transcription. IRF3 also drives intrinsic apoptotic pathways through interactions with Bax [7], and restrains NF-κB-driven inflammation through a separate, non-transcriptional inhibitory pathway [8].

Human IRF3 is a 427-residue protein with a two-domain architecture (Fig. 1.1). An N-terminal DNA-binding domain (DBD, residues 1–111) recognizes DNA se-

quences at the promoters and enhancers of IRF3 target genes, including the PRDIII-PRDI tandem element at the IFN-β enhancer and ISRE-like elements at directly IRF3-responsive genes [9]. A linker connects the DBD to a C-terminal regulatory domain (CTD, residues 181–427) and includes nuclear localization and export sequences. The CTD contains clusters of phosphoacceptor serines, the autoinhibitory C-terminal tail, the homodimerization interface, the binding surface engaged by upstream adaptor proteins, and the surface recognized by the transcriptional coactivators CBP and p300 [10, 11, 12, 13]. The activation mechanisms of IRF3 are therefore localized to IRF3CTD, and this domain is the focus of the IRF3-directed work presented in Chapters 2 and 3 of this dissertation.

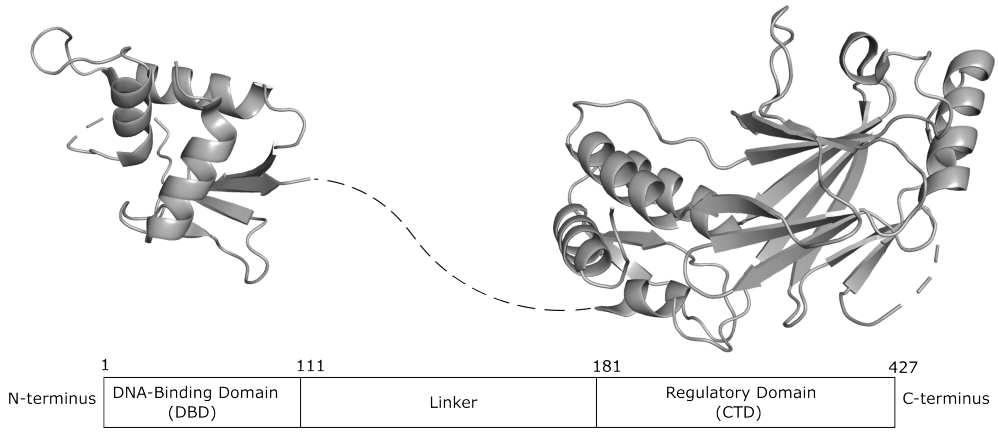


Figure 1.1: **Schematic diagram of IRF3.** Crystal structure of IRF3's DNA-binding domain (PDBID: 3QU6) and regulatory domain (PDBID: 1J2F), with a black, dashed line drawn to indicate the unstructured linker connecting the two domains.

1.2 The IRF3 activation cycle

IRF3 sits downstream of three principal pattern recognition receptor pathways, each recognizing viral nucleic acids. Although the upstream sensors and adaptors differ, all three pathways converge on a set of kinases that phosphorylate IRF3CTD.

TLR3-TRIF. Endosomal Toll-like receptor 3 (TLR3) recognizes viral double-stranded RNA generated during the replication cycles of RNA viruses. Upon

ligand binding, TLR3 recruits the adaptor TRIF (TIR-domain-containing adapter-inducing IFN-β) through homotypic TIR-domain interactions [2, 3]. TRIF in turn nucleates a signaling complex containing TRAF3 and the kinases TBK1 and IKKε, which carry out the activating phosphorylation of IRF3 [14, 15].

RIG-I/MDA5-MAVS. Cytosolic viral double-stranded RNA, generated during RNA virus replication, is sensed by the RIG-I-like receptors RIG-I (negative strand RNA viruses) and MDA5 (positive strand RNA viruses). Ligand-bound RIG-I and MDA5 engage the adaptor MAVS (mitochondrial antiviral signaling protein), which is anchored at the mitochondrial outer membrane through its C-terminal transmembrane domain, through CARD-mediated interactions [16, 17]. MAVS then recruits TRAF3, TBK1, and IKKε, again converging on the IRF3-activating kinase complex.

cGAS-STING. Cytosolic double-stranded DNA, derived from DNA virus genomes, reverse-transcribed retroviral intermediates, or mislocalized self-DNA, is detected by cyclic GMP–AMP synthase (cGAS). cGAS synthesizes the cyclic dinucleotide 2′3′-cGAMP, which binds and activates the endoplasmic-reticulum-resident adaptor STING (stimulator of interferon genes) [2, 3]. Activated STING traffics from the ER to the Golgi, where it recruits TBK1 and IRF3, and TBK1 phosphorylates STING-bound IRF3 [18].

The three pathways converge on TBK1 (and its close paralog IKKε), which phosphorylates a cluster of serines within the C-terminal "signal response" region of IRF3CTD. Systematic mutagenesis identified Ser-385, Ser-386, Ser-396, Ser-398, and Ser-402 as functionally relevant phosphoacceptors, with Ser-386 emerging as the minimal phosphorylation site required for activation [19, 20]. Phosphorylation at this cluster relieves the autoinhibited state of IRF3CTD and exposes the surfaces required for downstream protein–protein interactions.

The adaptor proteins that lead to the activation of IRF3 also require phosphorylation and share a common IRF3-binding motif. Liu et al. [13] demonstrated that MAVS, STING, and TRIF each contain a conserved pLxIS motif (p, hydrophilic residue; x, any residue; with the terminal Ser being the phosphoacceptor) at which they are phosphorylated by the same TBK1/IKKε kinases that

phosphorylate IRF3. The phosphorylated adaptors then recruit IRF3 through a positively-charged surface on IRF3CTD that recognizes the phospho-pLxIS sequence. IRF3CTD contains its own pLxIS region around Ser-396, and upon phosphorylation IRF3 engages the same positively-charged surface on a second IRF3 molecule to form the active homodimer [13, 12]. This unifying mechanism explains why the adaptor-binding surface and the dimerization surface on IRF3CTD overlap, and provides a structural basis for the convergence of multiple upstream pathways on IRF3. Notably, however, the phosphorylated IRF3 dimer crystallized by Jing et al. [12] carried a site-specifically incorporated phosphoserine only at Ser-386, with Ser-396 unmodified, indicating that phosphorylation at Ser-396 within the pLxIS motif is not strictly required for dimer formation *in vitro*. The precise sequence of phosphorylation events at the Ser-385/386 and Ser-396/398 clusters during physiological activation, and the relative roles of these two sites in adaptor recruitment versus dimer stabilization, remain incompletely resolved.

Following dimerization, IRF3 is recruited to the nucleus and engages the transcriptional coactivators CBP and p300 through a separate surface on IRF3CTD [21, 12]. The resulting IRF3-CBP/p300 complex assembles with NF-κB and ATF-2/c-Jun at the IFN-β enhanceosome to drive interferon transcription [22].

1.3 The Structure of IRF3CTD

The crystal structures of IRF3CTD published over the past two decades have made the activation cycle described above structurally interpretable, and they provide the structural context related to the fragment-screening results presented in Chapter 2.

The first crystal structures of IRF3CTD were reported independently by Qin et al. [10] and Takahasi et al. [11] in 2003. Both groups crystallized the unphosphorylated regulatory domain in the same space group and described an IRF-association core structurally homologous to the Smad MH2 domain, flanked by a C-terminal regulatory tail bearing the known phosphoacceptor sites (Fig. 1.2A). Despite the near-identical crystal forms, the two groups proposed distinct activa-

tion mechanisms. Takahasi et al. interpreted the dimer present in the asymmetric unit as a model of the activated state, in which the Ser-Ser-Leu motif of one subunit inserts into a hydrophobic pocket on the partner, with phosphorylated Ser-385/Ser-386 engaging a cluster of basic residues (R211, R213, K360, R361) at the rim of this pocket to stabilize the dimer. Qin et al. instead argued that the dimer present in the crystal structure is a crystallization artifact and proposed an autoinhibition-release model. They proposed that in the unphosphorylated form, the C-terminal tail is sequestered against a hydrophobic surface. Phosphorylation at serine residues 385, 386, 396, 398, and 402 destabilizes this autoinhibitory packing through charge repulsion, unmasking the hydrophobic surface for CBP binding.

A partial resolution to this discrepancy came with the structure of IRF3CTD bound to the IBiD domain of CBP [21]. CBP binds to a hydrophobic surface on IRF3CTD that overlaps with the surface engaged by the autoinhibitory C-terminal element in the latent structure, providing support for Qin et al.'s autoinhibition model: the coactivator-binding surface in the active state is the surface that is occluded in the latent state. The complete picture, however, required structures of the phosphorylated, active form.

In 2020, Jing et al. published crystal structures of the phosphorylated IRF3 homodimer in complex with CBP [12] (Fig. 1.2B). This structure showed the same association with CBP at IRF3CTD's hydrophobic surface that Qin et al. showed previously [21]. Jing et al.'s structure added to the picture that the C-terminal tail that dissociates from the hydrophobic surface wraps around to bind the partner subunit. Their crystal structure corroborated the earlier result from Servant et al., where a single phosphorylation at Ser-386 is sufficient for dimerization and activation. Jing et al. used genetic code expansion to incorporate phosphoserine at this single site and showed that this caused dimerization, where pSer-386 makes multiple contacts in a basic pocket in the partner subunit. Additionally, Jing et al.'s structure reveals that a portion of IRF3CTD's N-terminal tail (residues 189–199) packs against the CBP-binding interface in the inactive structure, dissociating in the active structure to allow for CBP binding. The results from these publica-

tions provide a coherent model of phosphorylation, dimerization, and coactivator recruitment.

In addition to these structures of IRF3CTD, two structures of IRF3 bound to viral antagonists have been reported. Rotavirus NSP1 binds IRF3CTD at a surface that overlaps with the adaptor- and dimer-binding region described above, mimicking the phospho-pLxIS recruitment mode used by MAVS, STING, and TRIF. While NSP1 can bind IRF3CTD without phosphorylation of its pLxIS motif, phosphorylation greatly enhances affinity [23]. The HPV16 E6 oncoprotein engages a distinct region of IRF3 DBD that overlaps with the linker (residues 102–113) [24]. Beyond these two cases, the binding modes of IRF3-targeting viral proteins are largely uncharacterized at atomic resolution. This gap motivates the biophysical work in Chapter 3.

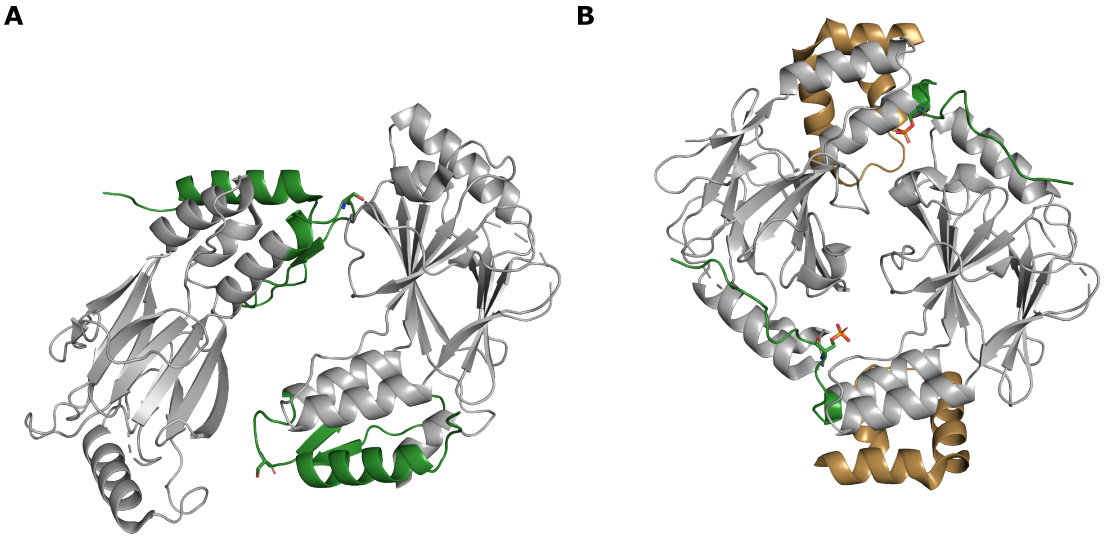


Figure 1.2: **Crystal structures reveal IRF3CTD's autoinhibitory mechanism. (A)** The unphosphorylated crystal structure of IRF3CTD published by Takahasi et al., with the C-terminal tail colored green (residues 381–422) and a stick representation shown for Ser386. (PDBID: 1J2F). The dimer seen here is a crystallization artifact. **(B)** Phosphorylated, biologically active dimer of IRF3CTD published by Jing et al., in complex with CBP (sand), with the C-terminal tail colored green (residues 381–396) and a stick representation shown for pSer386. (PDBID: 7JFL)

1.4 IRF3 in human disease

The central position of IRF3 in innate immune signaling makes aberrant IRF3 activity a contributor to a wide range of human pathologies, on both the loss-of-function and gain-of-function ends of the activity spectrum.

Loss of IRF3 function impairs antiviral immunity. A clear example of IRF3-related immune dysfunction came from a child carrying the IRF3 variant R285Q who presented with herpes simplex encephalitis (HSE) [25]. Functional analysis of the R285Q variant showed that it fails to undergo phosphorylation at Ser-386 in response to HSV-1 infection and is correspondingly deficient in IFN-β induction. The R285Q substitution sits in IRF3CTD and is one of the few human genetic variants in IRF3 with a clearly defined phenotype. More broadly, *Irf3*-deficient mice show heightened susceptibility to a wide range of viral infections, consistent with the role of IRF3 in establishing the rapid first-wave antiviral response [3].

Sustained or unregulated IRF3 activation drives inflammatory pathology. At the opposite end of the spectrum, persistent IRF3 activation is implicated in several "type I interferonopathies" and inflammatory conditions. STING-associated vasculopathy with onset in infancy (SAVI) is caused by gain-of-function mutations in STING that result in constitutive IRF3 activation. Affected children develop systemic vasculopathy and interstitial lung disease driven by sustained type I interferon signaling [3]. A subset of cases of Aicardi-Goutières syndrome, caused by mutations in nucleic-acid-processing enzymes that allow self-DNA to accumulate and engage cGAS-STING, similarly involve elevated IRF3-driven IFN-β expression [3]. Beyond interferonopathies, IRF3 has been shown to drive maladaptive inflammation in non-infectious settings. King et al. [26] demonstrated that ischemic cell death following myocardial infarction releases self-DNA that activates the cGAS-STING-IRF3 axis in cardiac macrophages, and that genetic deletion of *Irf3* protects mice from post-infarction inflammatory damage and mortality. IRF3 activation has also been linked to obesity-associated adipose inflammation and insulin resistance [27] and to sepsis-associated inflammatory injury [28].

IRF3's role in cancer is bidirectional. IRF3 contributes to anti-tumor

immunity through its role in the cGAS-STING response to tumor-derived DNA, and IRF3 activity correlates with improved outcomes in some tumor types [3]. However, IRF3 also contributes to disease progression in other contexts. In gastric cancer, for example, IRF3 has been identified as a positive regulator of tumor growth through its interaction with the Hippo-pathway transcriptional coactivator YAP, and pharmacological inhibition of IRF3-YAP signaling has been proposed as a therapeutic strategy [29].

IRF3 is therefore an attractive therapeutic target across multiple indications, with the direction of modulation depending on disease context [3, 30, 31]. Antiviral and vaccine-adjuvant applications motivate the development of IRF3 activators or pathway agonists, while interferonopathies, post-infarction injury, certain cancers, and inflammatory conditions motivate the development of IRF3 inhibitors.

1.5 Viral antagonism of IRF3

Because IRF3 is critical for a rapid antiviral response, many viral pathogens have evolved mechanisms for suppressing IRF3 activation [3, 30]. Viral antagonists target IRF3 at every step of its activation cycle. Upstream, viral proteases can sequester or degrade the sensors and adaptors that initiate signaling. This is exemplified by hepatitis C virus NS3/4A cleavage of MAVS and TRIF. At the kinase level, viral proteins can disrupt the TBK1/IKKε complex, as seen with SARS-CoV M protein and encephalomyocarditis virus 3C protease. At the level of IRF3 itself, some antagonists block IRF3 phosphorylation directly (porcine reproductive and respiratory syndrome virus NSP1β, influenza A NS1). Once IRF3 is phosphorylated, viral proteins can promote its degradation (HSV ICP0, varicella-zoster ORF61, HIV Vpu) or block IRF3 dimerization or nuclear translocation (Theiler's virus leader protein, West Nile NS1). Even nuclear, transcriptionally active IRF3 is targeted (Zika NS5). Finally, some viruses encode decoy mimics of IRF3, such as HHV8 vIRF-3. These mechanisms are catalogued in detail in recent reviews [3, 30].

Despite the breadth of cellular and genetic evidence for IRF3 antagonism by

viral proteins, the biophysical and structural characterization of these interactions is sparse. As noted above, only two viral antagonists, rotavirus NSP1 [23] and HPV16 E6 [24], have been structurally characterized in complex with IRF3. For most other reported antagonists, the molecular evidence consists of co-immunoprecipitation from cell lysates, co-localization microscopy, and functional reporter assays, with little or no validation that the proposed interaction is direct or specific in a purified-component setting. This biophysical-characterization gap is consequential because the cellular environment contains many opportunities for indirect association through shared scaffolding partners, induced post-translational modifications, or co-recruitment to membranes or organelles that can be mistaken for direct binding. Establishing whether a proposed viral antagonist binds IRF3 directly, and if so where and with what affinity, is a prerequisite for both mechanistic understanding and structure-guided intervention.

Respiratory syncytial virus (RSV) nonstructural protein 1 (NS1) provides a representative example of this gap and is the focus of Chapter 3. Spann et al. [32] showed that NS1 and NS2 cooperatively suppress IRF3-dependent gene induction in RSV-infected cells. A subsequent study by Ren et al. [33] proposed a specific mechanism. NS1 was reported to associate with both IRF3 and CBP in co-immunoprecipitation experiments, and to disrupt the IRF3-CBP interaction without blocking the upstream phosphorylation, dimerization, or nuclear translocation of IRF3. This model is mechanistically distinct from most other reported viral antagonism strategies and would be of substantial therapeutic interest if confirmed at the biophysical level. To date, however, no direct measurement of NS1 binding to IRF3 with purified components has been reported. Chapter 3 describes our attempt to characterize this proposed interaction *in vitro*.

1.6 Chemical tractability of IRF3 and crystallographic fragment screening

Despite broad therapeutic interest in IRF3, the field has lacked experimentally validated small-molecule ligands of the protein itself. This is consistent with the

general difficulty of targeting transcription factors with small molecules. Transcription factors typically lack the deep, well-defined pockets that have historically been the most tractable targets for medicinal chemistry. Additionally, their key functions involve interacting with other proteins and DNA, as opposed to the more easily targeted active site chemistry in enzymes [31]. Reported chemical modulators of the IRF3 pathway to date all act upstream of IRF3 itself: the small-molecule IRF3 agonist reported by Probst et al. [34] acts through the RIG-I-like receptor (RLR) pathway, the molecules identified by Foss et al. [35] similarly engage upstream components, and the carotenoid fucoxanthin has been proposed to bind IRF3 directly based on docking calculations [28], but without experimental validation of binding. Direct chemical matter for IRF3 is therefore essentially absent from the literature.

Crystallographic fragment screening is a powerful method for identifying chemical matter for protein targets that have resisted conventional approaches [36, 37]. In a typical campaign, crystals of the target protein are soaked individually with a library of several hundred to several thousand low-molecular-weight, chemically diverse fragments and X-ray diffraction data are collected from each soaked crystal. A major advantage of this approach is the relative simplicity of the fragments compared to the larger molecules required for detection in conventional high-throughput assays. The fragment's lower molecular complexity gives each fragment a higher per-molecule probability of binding any given site, so a modest library covers chemically diverse interaction features more efficiently than an high-throughput screening library of drug-like compounds.

However, individual fragments bind weakly and with low occupancy, so conventional difference-map analysis is insufficient to detect them reliably. The PanDataset Density Analysis (PanDDA) algorithm [38, 39] addresses this problem by leveraging the large number of essentially isomorphous data sets generated in a fragment campaign. Ground-state electron density is estimated by averaging across many apo (or apo-equivalent) data sets, and fragment-bound data sets are then identified as outliers from this ground state. PanDDA event maps reveal low-occupancy binders that would not be interpretable, or even detectable, in

conventional difference density maps. The resulting fragment hits provide both chemical matter and structural information about the protein surface, including the identification of pockets that may not have been apparent in apo structures. Fragment-screening campaigns have been used successfully to identify chemical matter on viral protein targets [40, 41], on allosteric pockets of disease-relevant enzymes [42, 43], and to identify novel protein-protein interaction sites [44]. Chapter 2 applies this approach to IRF3CTD.

1.7 Conformational dynamics and the redesign of enzyme function

Chapter 4 of this dissertation presents two collaborative projects with the Chica group at the University of Ottawa in which crystallographic information about protein conformational ensembles is used to improve and validate the computational design of enzymes. Enzymes are dynamic, flexible macromolecules that sample multiple conformational states, the relative populations of which determine function. Computational protein design has historically been formulated as a single-state problem, in which a fixed backbone template is provided and amino-acid sequences are optimized for compatibility with that template. Single-state design has been successful for improving stability, redesigning binding interfaces, and grafting catalytic activities into new scaffolds, but it is poorly suited to problems whose outcome depends on the relative stability of multiple conformational states or on the population of catalytically productive substates within a single state [45]. Crystallographic methods that model the conformational ensemble of a protein, rather than a single structure, address this limitation in two ways. First, the ensembles serve as inputs to design procedures that account for conformational heterogeneity. Second, the same methods can be applied after the fact to verify that engineered mutations produced the conformational shifts they were intended to produce. The following design strategies developed in the Chica laboratory illustrate these two roles in distinct enzymatic systems.

Multistate design and the AAT conformational landscape. The first

strategy, multistate computational design, scores candidate sequences against multiple backbone templates simultaneously and selects those that preferentially stabilize one state over another [45]. By computing the energy of each candidate sequence on each target state, multistate design can be used to remodel a protein's conformational landscape rather than only its static structure. The model system to which the Chica laboratory applied this approach, and the focus of the first project in Chapter 4, is *E. coli* aspartate aminotransferase (AAT), a pyridoxal-phosphate-dependent enzyme that catalyzes the transamination of L-aspartate with α-ketoglutarate. AAT provides a well-studied example of conformational exchange that influences catalysis. AAT undergoes hinge-mediated domain motions during its catalytic cycle, transitioning between an open conformation accessible to substrate and a closed conformation in which the active-site residues are organized for catalysis [46]. The enzyme adopts an open conformation in the apo state, and substrate binding drives closure of a small domain over the active site through rotation about a hinge. The open/closed equilibrium is a determinant of both catalytic efficiency and substrate selectivity because access to the catalytically competent closed state depends on the binding energy contributed by the substrate.

The role of this conformational equilibrium in AAT substrate selectivity was described by Onuffer and Kirsch, who substituted six residues of *E. coli* AAT with the corresponding residues from the homologous *E. coli* tyrosine aminotransferase [46]. The resulting hexamutant (HEX) acquired three orders of magnitude higher transamination activity with the non-native substrate L-phenylalanine without sacrificing activity on the native substrate L-aspartate. A subsequent crystallographic study by Malashkevich et al. revealed that the HEX enzyme adopts the closed conformation in its ligand-free form, in contrast to wild-type AAT, which is open in the absence of ligand [47]. The HEX mutations therefore did not simply remodel the active site to accommodate aromatic substrates. Rather, they shifted the conformational equilibrium of the enzyme toward the closed state, pre-organizing the active site for catalysis on substrates whose binding energy is too low to drive closure on their own. The observation that altered substrate selectivity can be

encoded in a shifted conformational equilibrium, rather than in altered active-site geometry alone, frames the design problem addressed in the AAT project of Chapter 4. That project applies the Chica laboratory's multistate-design strategy to AAT, using crystallographic models of the open and closed states as input templates, with the goal of designing variants that preferentially adopt the closed or the open conformation. Changes in catalytic specificity and efficiency between the two states served as the key functional benchmark for design success.

The AAT project was performed in collaboration with the Chica group, who developed and executed the multistate-design strategy and the kinetic characterization of the designed variants. My contribution focused on the multi-temperature X-ray crystallography of the designed variants, providing structural validation that the predicted conformational shifts had occurred.

Ensemble-based design and the Kemp eliminases. The second strategy developed in the Chica laboratory uses crystallographic models in a different way, not to score sequences against multiple well-defined functional states but to better sample the conformational diversity within a single state. Crystallographic ensemble refinement produces a set of backbone conformations that together fit the diffraction data, providing a representation of heterogeneity that single-model refinement averages out. Active-site sequences are then designed against each member of the ensemble rather than against a single fixed template, allowing the design procedure to identify substates that are catalytically productive but poorly represented in the average structure. Applied to the *de novo* designed Kemp eliminases HG3 and KE70, this approach yielded variants with 100- to 250-fold improvements in catalytic efficiency, comparable to those reached by extensive and costly directed-evolution campaigns. Room-temperature crystal structures of the most active variants showed both higher complementarity of the active site to the transition state and a narrower conformational ensemble than the parent enzymes, indicating that the activity gains came from enrichment of catalytically competent substates rather than from active-site rebuilding alone.

The Kemp eliminase project was likewise performed in collaboration with the Chica group, who developed the ensemble-based design strategy and performed

the kinetic characterization of the designed variants. My contribution was the collection of room-temperature X-ray diffraction data confirming the predicted designs.

The two strategies thus address related but distinct questions about conformational design, with multistate design focused on the relative populations of well-defined functional states and ensemble-based design focused on the enrichment of productive substates within a single state.

1.8 Dissertation overview

The following three chapters detail the results of the projects mentioned above. **Chapter 2** presents a crystallographic fragment screen of IRF3CTD, the first such campaign on this protein, which identified 25 fragment hits distributed across eight distinct binding sites on the domain. Three of these sites are positioned adjacent to functional surfaces: the autoinhibitory C-terminal tail, the dimerization and adaptor-binding hot spot, and the CBP coactivator interface. These results provide the first experimentally validated starting points for direct chemical intervention on IRF3. **Chapter 3** describes an *in vitro* biophysical investigation of the proposed interaction between RSV NS1 and IRF3. Using size-exclusion chromatography, fluorescence polarization, and biolayer interferometry with purified components, we tested for direct binding using both unphosphorylated, monomeric IRF3 and phosphorylated, dimeric IRF3. Specific binding was not observed. We discuss the implications of this result for the proposed NS1-IRF3-CBP disruption model and the methodological lessons for biophysical validation of host–pathogen interactions more broadly. **Chapter 4** presents two collaborative projects with the Chica group. The first is a multistate-design remodelling of the AAT conformational landscape, in which preferential stabilization of the catalytically active closed state is used to alter substrate selectivity toward the non-native L-phenylalanine substrate. Designed variants are characterized by steady-state kinetics, room-temperature X-ray crystallography, and multitemperature ^{19}F NMR, providing direct structural and dynamic evidence that the engineered mutations remodel the

open/closed equilibrium as intended. The second is an ensemble-based redesign of the *de novo* Kemp eliminases HG3 and KE70 that increases their catalytic efficiency, with room-temperature crystallography confirming that the activity gains arise from enrichment of catalytically competent substates.

Chapter 2

Crystallographic fragment screen of IRF3CTD

2.1 Introduction

Interferon regulatory factor 3 (IRF3) is a transcription factor that activates the innate immune system in response to viral infection [1]. Upon detection of pathogen-associated molecular patterns, multiple signaling pathways converge on the C-terminal regulatory domain of IRF3 (IRF3CTD), where phosphorylation by the kinases TBK1 and IKKε [15] relieves an autoinhibited state, allowing IRF3 to homodimerize and translocate to the nucleus [48, 2]. IRF3's N-terminal, DNA-binding domain (IRF3DBD) then binds various promoters to drive transcription of interferons and other antiviral effectors. IRF3CTD mediates the principal protein–protein interactions that govern IRF3's activation cycle. It contains the cluster of phosphoacceptor serines [19, 20], the autoinhibitory C-terminal tail, and the dimerization interface [49]. It also includes the conserved binding site of adaptor proteins that mediate phosphorylation, MAVS, STING, and TRIF, as well as the surface that engages the transcriptional co-activator CBP [13, 23, 49].

Beyond this canonical role in interferon induction, IRF3 also participates in non-transcriptional pathways. It is a key effector of the RIG-I-like receptor-induced IRF3-mediated pathway of apoptosis (RIPA), where Bax binds IRF3CTD to ini-

tiate apoptosis in response to viral infection [7]. IRF3 also inhibits the nuclear translocation of NF-κB to restrain inflammation (RIKA) [8], where IRF3CTD binds the p65 subunit of NF-κB.

Aberrant IRF3 activity is implicated in human disease. Loss of IRF3 function impairs antiviral immunity, while sustained or excessive IRF3 activation drives pathological inflammation in conditions including post-infarction myocardial damage [26], adipose inflammation and insulin resistance [27], and sepsis [28]. IRF3 has also been shown to associate with YAP to drive the transcription of oncogenes in gastric cancer [29]. Additionally, many viruses encode antagonists of IRF3 to disable the host interferon response, including SARS-CoV-2 spike [50], HIV Vpu [51, 52, 53], and RSV NS1 [32]. IRF3 is therefore an attractive therapeutic target for both antiviral and immunomodulatory indications [3, 30].

As is the case with most transcription factors, which typically lack the well-defined pockets that make enzymes tractable drug targets, IRF3 has historically been considered a difficult target for direct chemical intervention. Experimentally validated ligands for IRF3 have not been published. Reported modulation of the IRF3 pathway has so far been limited to indirect agents such as the small-molecule agonists identified by Probst et al. [34] and Foss et al. [35], all of which function upstream of IRF3, and to a compound that dampens the IRF3-dependent inflammatory response, but is only proposed to act on IRF3 through a docking result [28].

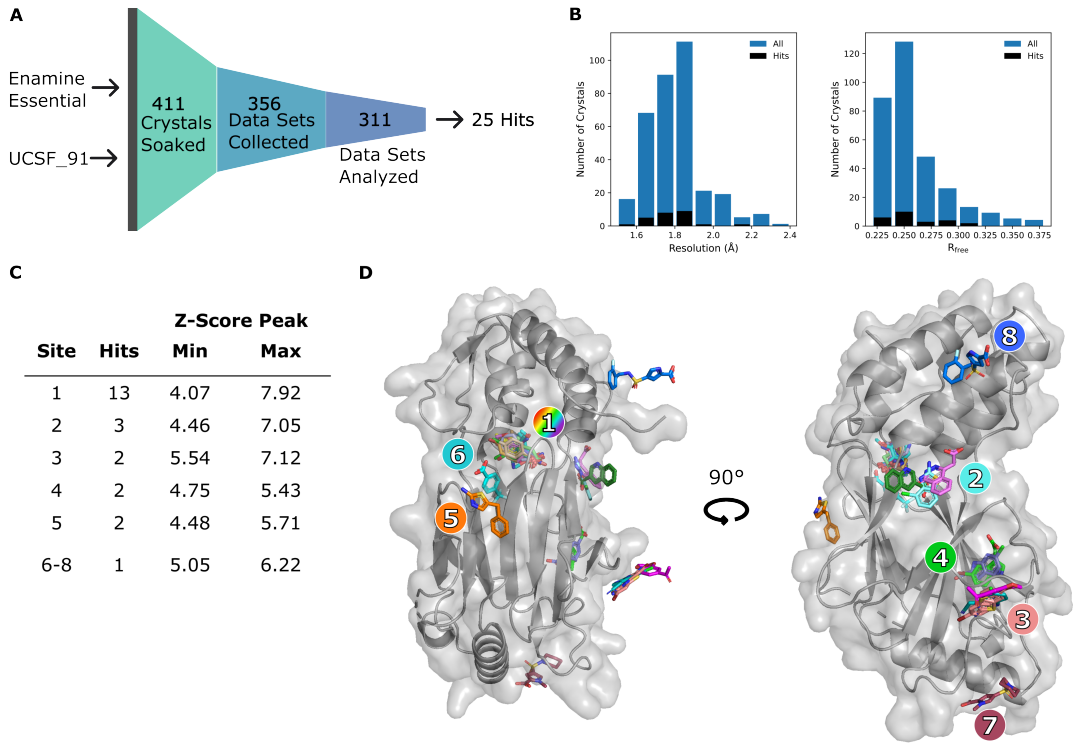
Crystallographic fragment screening provides an efficient route to identifying chemical matter for previously unliganded targets such as IRF3, mapping the protein surface with chemically diverse, low-molecular-weight probes whose binding poses are resolved at atomic detail [36, 37]. Here, we report a crystallographic fragment screen of IRF3CTD that yielded 25 fragment hits distributed across eight distinct binding sites (Fig. 2.1). These hits provide the first set of chemical starting points for targeting IRF3, and the structural information they provide will support the development of compounds that modulate IRF3 activation, probe its interactions with adaptors and viral antagonists, or serve as tools for further mechanistic investigation.

2.2 Results

IRF3CTD crystals were grown as previously reported [11], with the addition of seeding to achieve consistent growth in large batches. The crystals were soaked with 320 fragments from Enamine's Essential Fragment Library and 91 fragments from an in-house library (UCSF_91), for a total of 411 fragments. Of these, 311 crystals yielded diffraction data of sufficient quality for fragment-hit identification with the Pan-Dataset Density Analysis (PanDDA) method (Fig. 2.1B) [38]. PanDDA analysis identified 28 data sets with hits, although ligands could not be modelled for three of the data sets. Ligands for the remaining 25 data sets were modelled and statistically validated (see Methods). The hits are distributed across eight distinct binding sites on IRF3CTD (Fig. 2.1C). The three most functionally significant sites, 1, 2, and 7, are described below. These three sites are of particular interest because each sits adjacent to the autoinhibitory C-terminal tail or the dimerization/adaptor-binding surface that controls IRF3 activation.

2.2.1 Site 1 includes 13 hits in a buried hydrophobic pocket

Site 1 is a buried pocket located between the central β-sheet of IRF3CTD and the C-terminal bundle of helices (Fig. 2.2A). In the unbound structure, the pocket is occluded by a π-stacking interaction between W202 and R276 and is otherwise lined by tightly packed hydrophobic residues (Fig. 2.2B, sand). Fragment binding at this site results in the rearrangement of W202 and F204 to open the pocket, allowing the aromatic or hydrophobic ring of the bound fragment to pack between R276 and C371. Across all 13 hits, an oxygen atom of the fragment, most often part of a carboxyl group, anchors the fragment to the backbone carbonyl of W280 or Q279. At the crystallization pH of 5.1, only slightly above the typical pK_a range of aromatic carboxylic acids, a fraction of the carboxyl group population is expected to be protonated and competent to donate a hydrogen bond to the backbone carbonyl. Van der Waals contacts with the surrounding hydrophobic residues stabilize each fragment to varying degrees. The 13 fragments generally share a benzoic acid scaffold, differing primarily in the substitution pattern of



Site	Hits	Z-Score Peak Min	Z-Score Peak Max
1	13	4.07	7.92
2	3	4.46	7.05
3	2	5.54	7.12
4	2	4.75	5.43
5	2	4.48	5.71
6-8	1	5.05	6.22

Figure 2.1: **Crystallographic fragment screen of IRF3CTD yields 25 hits across eight binding sites.** **(A)** Overview of the crystallographic fragment screening campaign. **(B)** Resolution and R_{free} distributions for the 311 data sets used for PanDDA analysis. **(C)** Summary of fragment hits identified at each binding site, with the range of Z-score peaks observed across hits at each site. Z-scores indicate the significance of the fragment-bound density relative to the mean density of unbound ground-state. **(D)** Overview of fragment hits, with fragments shown as sticks and sites labeled 1–8. Fragment hits span the surface of IRF3CTD and include a buried hydrophobic pocket with 13 hits (site 1).

the aromatic ring (Fig. 2.2C). Some fragments deviate from this scaffold with a slightly different ring structure (Fig. 2.2D) or a slightly different polar group structure (Fig. 2.2E).

2.2.2 Three fragments bind site 2, a surface pocket adjacent to the C-terminal tail

Site 2 is a shallow surface pocket adjacent to the IRF3CTD C-terminal tail (Fig. 2.3A). The pocket is lined by D238, W280, F320, and D321, and offers

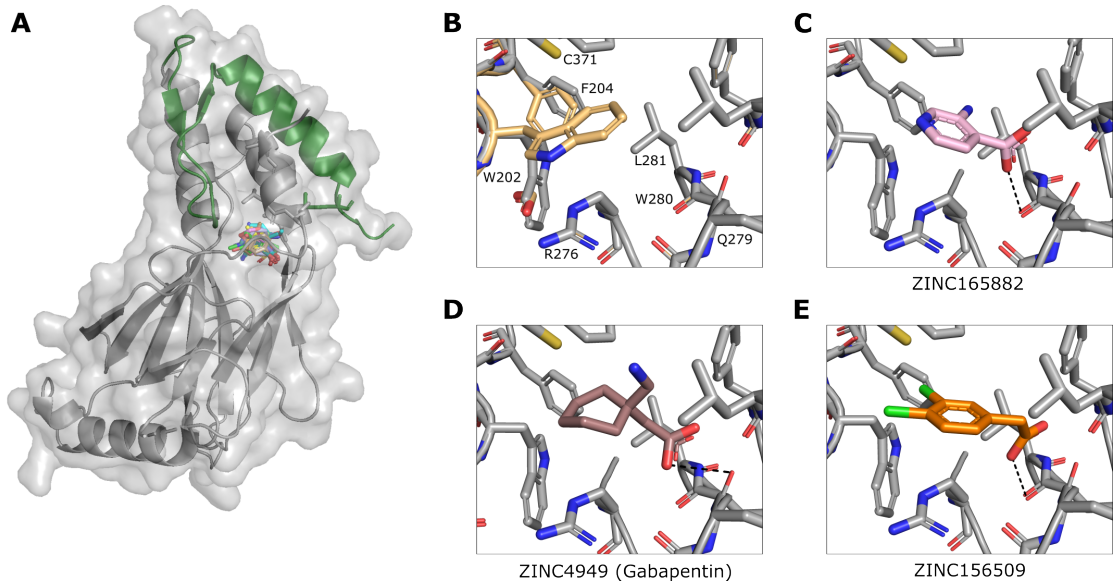


Figure 2.2: **Site 1. (A)** Overview of 13 fragments bound in site 1, shown as sticks within the hydrophobic core of IRF3CTD. The C-terminal tail involved in homodimerization is shown in green. **(B)** Unbound (sand) and fragment-bound (gray) structures of the site 1 pocket, highlighting the W202 and F204 rearrangement that accompanies fragment binding. **(C–E)** Representative fragment hits in site 1. Dashed lines indicate the typical hydrogen bond between a fragment oxygen and a backbone carbonyl of W280 or Q279.

polar anchors at the D238 side chain and along the V319-D321 backbone. The three fragments that bind here engage different combinations of these residues. ZINC8615114 hydrogen-bonds primarily to the backbone oxygen and nitrogen of D321 and makes additional weak van der Waals contacts with W280 and F320 (Fig. 2.3B). ZINC8652361 instead anchors to the backbone carbonyl of V319 and packs against the side chains of W280 and D321 (Fig. 2.3C). ZINC265642 spans the pocket, hydrogen-bonding both to the D238 side chain and to the D321 backbone while making van der Waals contacts with F320 (Fig. 2.3D). Although none of the individual fragments engage all of the pocket's polar anchors simultaneously, their complementary binding poses together map the chemical and geometric features that an elaborated ligand might exploit.

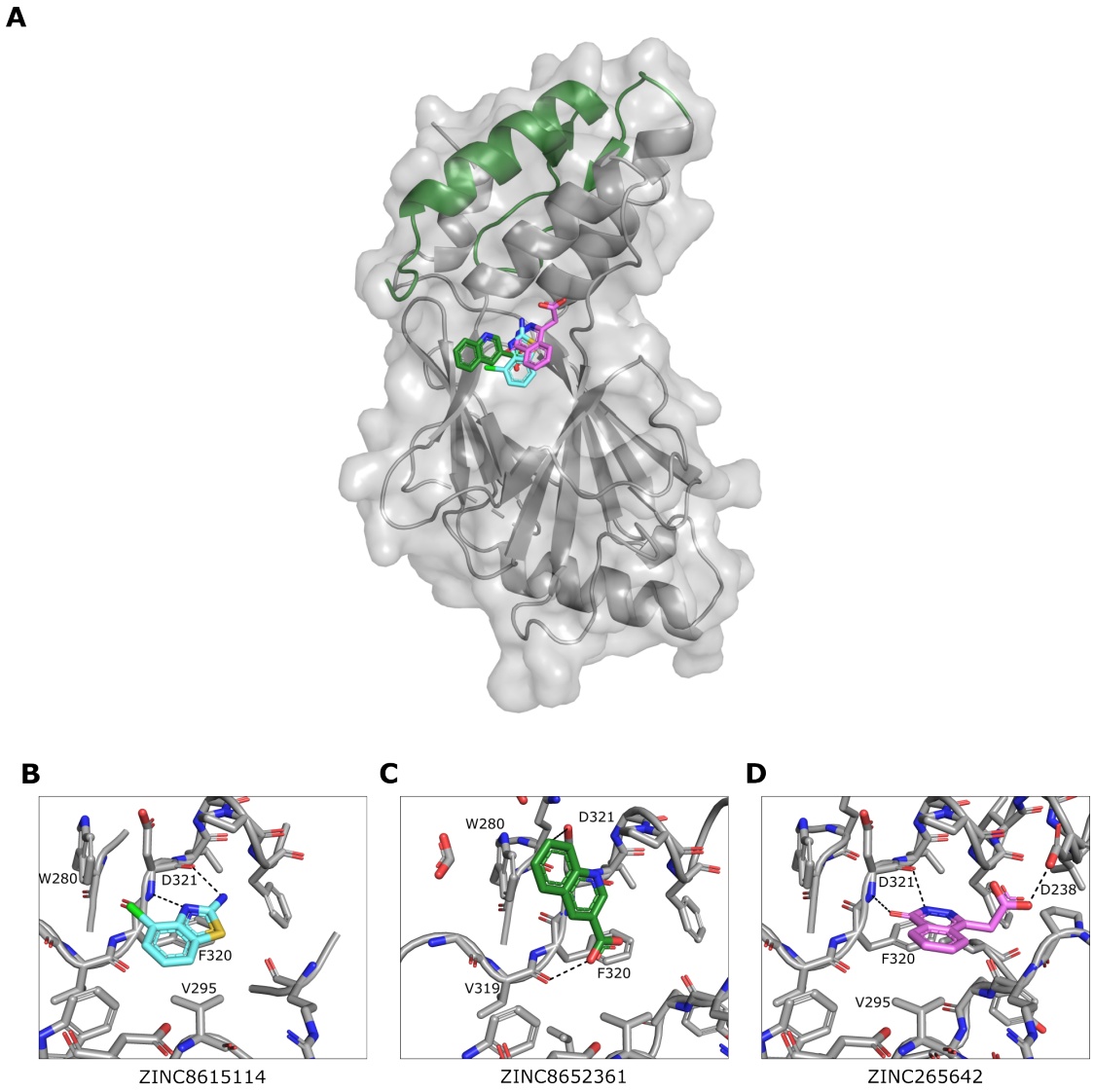


Figure 2.3: **Three fragments bind site 2, adjacent to the C-terminal tail. (A)** Site 2 overview, with fragment hits shown as sticks and the IRF3CTD C-terminal tail shown in green. **(B–D)** Fragment hits in the site 2 pocket, with key contacts shown as dashed lines.

2.2.3 One fragment binds at a key dimer and adaptor interface

Site 7 is a pocket formed between a loop and an α-helix on the surface of IRF3CTD that overlaps with the binding sites of multiple biologically important partners (Fig. 2.4A,B). This site anchors the C-terminal tail of the activated IRF3

homodimer [12] and is also the site at which the adaptor proteins MAVS, TRIF, and STING engage IRF3CTD during activation [23]. Notably, the same surface is targeted by rotavirus NSP1, which exploits this binding mode to antagonize the host interferon response [23]. One fragment from our screen, ZINC20283605, binds in this pocket (Fig. 2.4C). It forms polar contacts with H263 and E350 and packs against the backbone of G256 and the side chains of V257, L362, and K360. The convergence of many physiological and pathogen-derived partners on this single surface, combined with the availability of a fragment binder, makes site 7 a particularly attractive starting point for chemical-probe development.

2.3 Discussion

IRF3 sits at the convergence point of multiple innate immune signaling pathways and, through its non-transcriptional roles in apoptosis and inflammation, contributes to a broad spectrum of human diseases [3, 30]. Inhibition of IRF3 has been proposed as a strategy for limiting inflammatory pathology in conditions such as sepsis [28] and post-infarction myocardial injury [26], while activation or stabilization of IRF3 is of interest for antiviral and vaccine-adjuvant applications [34]. IRF3CTD is targeted by many viruses to disable the host interferon response [54, 32, 55, 33, 56, 25, 57, 58, 59, 60, 24, 61]. Although the binding sites of most of these antagonists on IRF3 are not yet resolved, rotavirus NSP1 has been shown to bind IRF3CTD's adaptor protein hot spot in a crystal structure (Fig. 2.4B) [23]. Considering IRF3CTD's role in transcription regulation and that it contains many known protein-protein interaction sites, it is plausible that this domain is commonly targeted by viral proteins. Therefore, ligands that bind this domain are valuable tools for investigating viral protein-IRF3 interactions and may serve as therapeutics. Additionally, if multiple viruses target a common protein-protein binding site on IRF3CTD, such as the site targeted by rotavirus NSP1, a drug that blocks that site from viral protein binding may serve as a broad-spectrum therapeutic. The fragment hits presented here provide the first experimentally validated chemical starting points for such efforts.

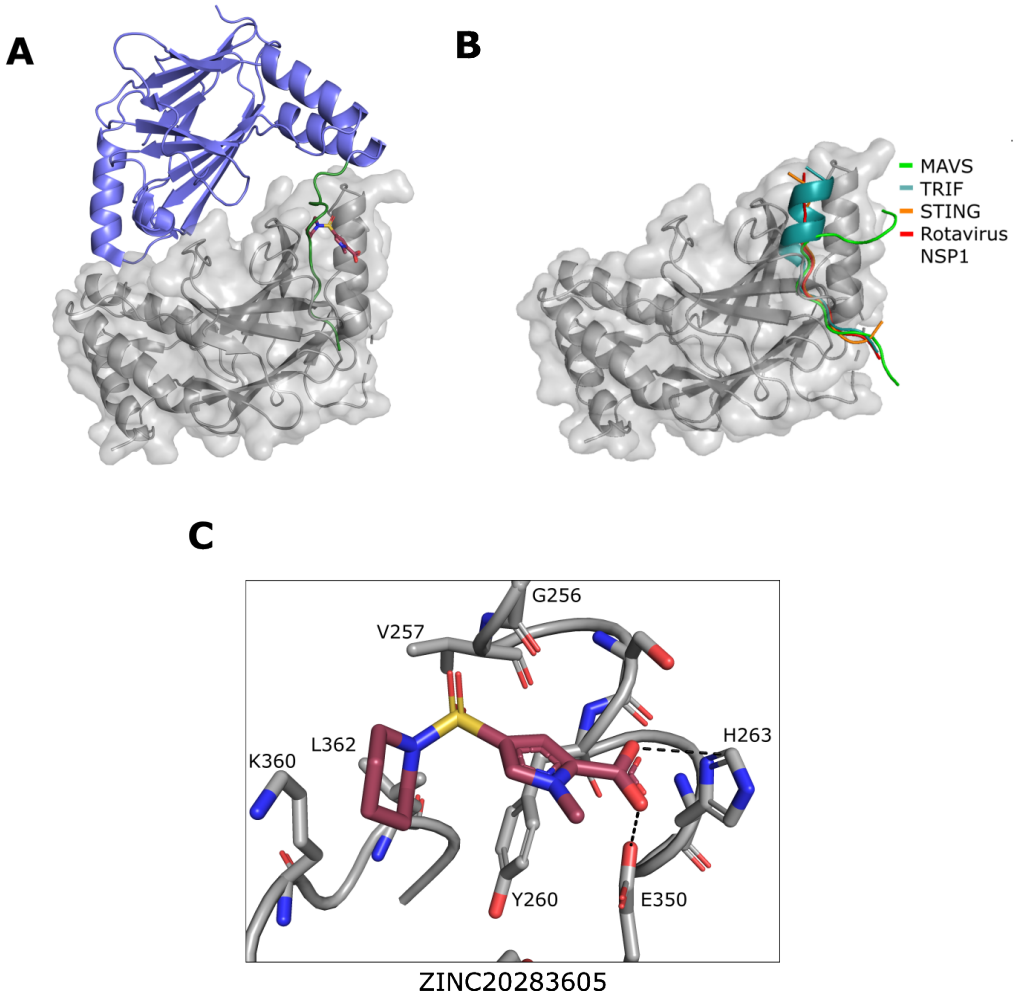


Figure 2.4: **Site 7. (A)** Overview of site 7, aligned with one chain from the crystal structure of the active IRF3 homodimer shown in slate, with the C-terminal tail shown in green (PDBID: 7JFL). The fragment hit is shown as sticks. **(B)** Aligned crystal structures of proteins that bind IRF3CTD at site 7: MAVS (5JEK), TRIF (5JEL), STING (5JEJ), and rotavirus NSP1 (5JEO). **(C)** Key interactions between ZINC20283605 and the site 7 pocket.

Site 1, with 13 hits, sits in IRF3CTD's hydrophobic core. This core partially stabilizes the autoinhibited state of IRF3CTD [10] and one face of this pocket becomes exposed in the phosphorylated, active conformation. Both C-terminal and N-terminal tails of IRF3CTD bind to one face of this core and dissociate from it to expose the hydrophobic surface that CBP binds in order to form the transcriptionally active complex [12] (Fig. 2.5A,B). The positions of the residues that anchor fragment binding at this site, including the potential covalent binding

target C371, are preserved between the active and inactive states, suggesting that our fragments could engage IRF3CTD in either conformation. Superposition of site 1 fragments onto the active, CBP-bound structure places the fragment between 3 and 4 Å of CBP residue I2089, indicating that an elaborated compound at this site could plausibly interfere with activation and CBP recruitment (Fig. 2.5B).

Site 2 is located on IRF3CTD's surface and corresponds to the region where the carotenoid fucoxanthin has been predicted to bind IRF3 in docking studies [28]. Su et al. report that fucoxanthin inhibits IRF3 phosphorylation in models of sepsis. Given that this pocket is not located at the IRF3 phosphoacceptor site, such an effect may be explained by interference with the scaffold interactions required for phosphorylation or other allosteric mechanisms, rather than by a direct effect on the phosphorylation chemistry. The three fragments bound in site 2 engage distinct combinations of the pocket's polar anchors and may be used to validate fucoxanthin binding through competition assays, or may themselves interfere with phosphorylation. Additionally, site 2 is adjacent to the C-terminus, which is likely flexible and could not be resolved in our crystal structures. However, the side chain of Q422 may be positioned as close as 4 Å to ZINC8652361 (Fig. 2.5C). This raises the possibility of developing a compound that binds site 2 and anchors the C-terminal tail to the main body of IRF3CTD, potentially preventing activation.

The fragment bound in site 7 may have the potential for probing protein-protein interactions. As shown in Fig. 2.4A,B, site 7 is a hot spot for protein–protein interactions on IRF3CTD. The activated dimer's C-terminal tail anchors here and the adaptor proteins STING, MAVS, and TRIF all engage this surface during IRF3 recruitment and activation [13, 23]. The same site is targeted by rotavirus NSP1 as part of its strategy for suppressing host interferon induction [23]. Interestingly, while these proteins bind many residues across the adjacent α-helix and random coil, H263 is a common anchor. This is also a main interaction for ZINC20283605, suggesting that this fragment may be able to block interactions between the adaptor proteins/IRF3 tail and H263. This fragment is, therefore, a particularly valuable starting point: an elaborated compound at site 7 could in principle interfere with adaptor recruitment, dimer stability, or viral antagonism.

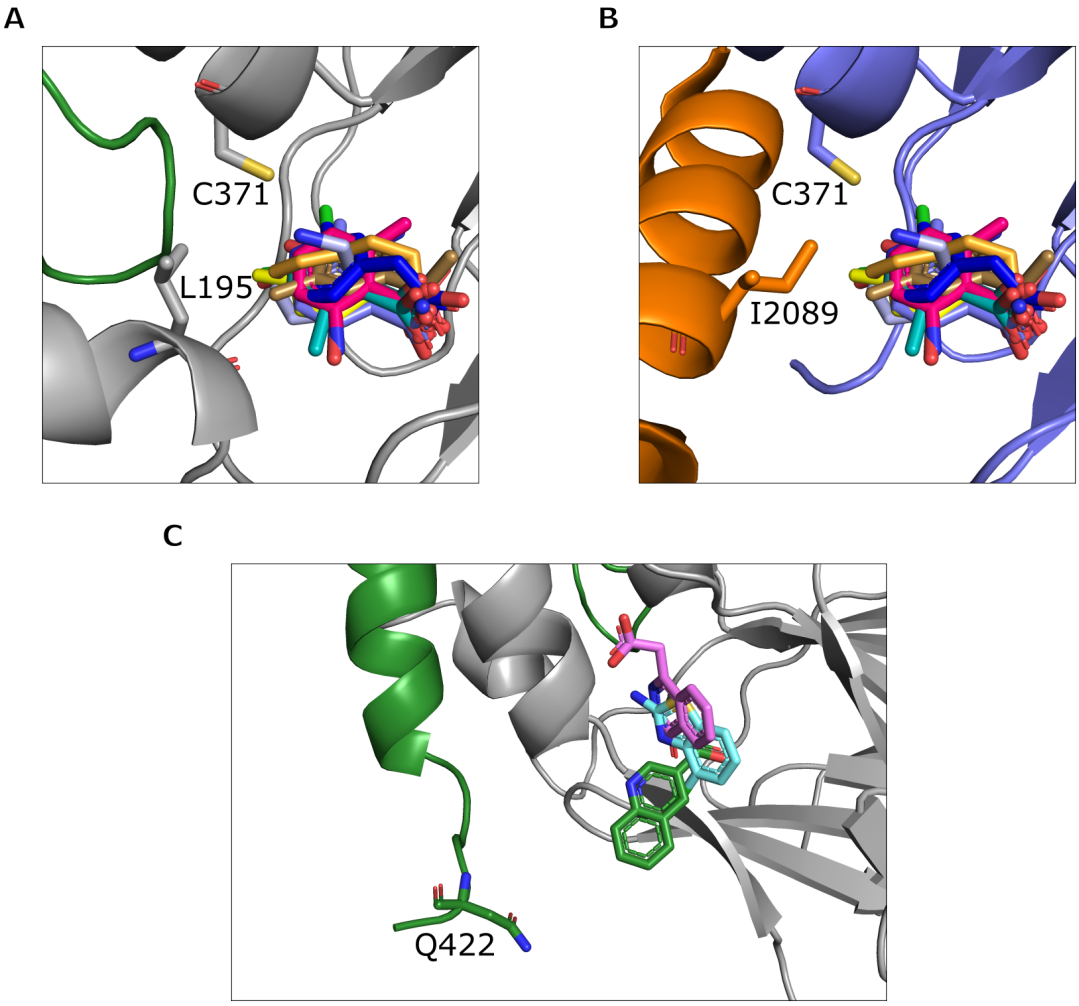


Figure 2.5: **Sites 1 and 2 contain fragments in close proximity to key interfaces.** **(A)** Site 1 fragments (colored sticks) make contact with the N-terminal residue L195 and are adjacent to the C-terminal tail, shown in green. **(B)** Site 1 fragments were aligned to the crystal structure of phosphorylated IRF3CTD (slate) bound to CBP (orange) (PDBID: 7JFL). Superimposed in this way, they make contact with I2089 of CBP. **(C)** Site 2 fragments bind near the C-terminal tail (green). Q422 is shown as sticks. Because no electron density was visible for this side chain, a predicted side chain rotamer was chosen from a rotamer library based on minimizing clashes with the backbone.

2.4 Fragment Development

Crystallographic fragment screening is a highly sensitive method for detecting fragments, typically resulting in much higher hit rates than traditional high-throughput screening. The tradeoff for this sensitivity is that many detected fragments bind very weakly, often failing to be detected in solution binding experiments. Thus, fragment development is typically required to move from hits to active compounds. Linking adjacent fragments or merging fragments that partially overlap with one another are effective strategies to improve binding affinity.

The site 1 fragments share a common binding pose and represent only modest chemical diversity around a benzoic-acid scaffold, which limits the prospects for conventional fragment linking or merging within this site. Additionally, the hydrogen bond between the fragment carboxyl group and the backbone carbonyl will likely not form at physioliogcal pH, where the cayboxyl group will be entirely deprotonated. This brings into question the practical potential of these fragments to bind site 1 in a biological context. Two potential modifications could be made to these fragments to overcome these challenges. First, the carboxyl group could be replaced with a similar group that can donate a hydrogen bond at physiological pH, such as a primary amide. Second, the proximity of C371 to the bound fragments may enable covalent covalent binding (Fig. 2.6A,B). Simple electrophilic substituents such as nitriles or aldehydes may be added the aromatic ring to engage the cysteine residue, as has been shown in a number of successful drugs [62].

Sites 2, 3, and 4 each contain multiple hits in distinct but overlapping orientations, and so are amenable to fragment linking or merging strategies. As an example, Fragmenstein [63] (see Methods) predicted an improved binding free energy for a merged molecule derived from site 2 fragments (Fig. 2.6C). Compounds developed along these lines would not be expected to reach drug-like potency without further optimization, but they would provide valuable tool compounds for solution-based characterization of IRF3 binding and for testing hypotheses about the functional consequences of ligand occupancy at each site.

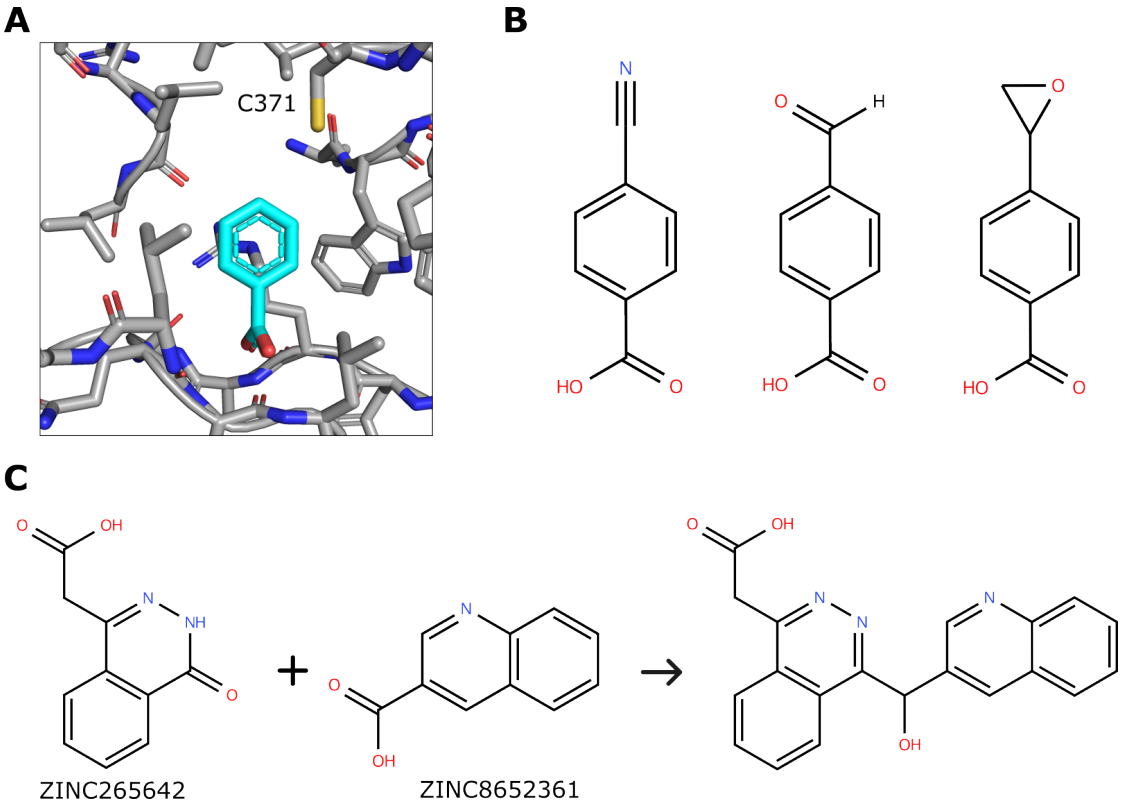


Figure 2.6: **Potential routes for fragment development.** **(A)** Hypothetical scaffold (cyan) placed in site 1, adjacent to C371. **(B)** Candidate cysteine-reactive warheads that could engage C371. **(C)** Fragmenstein-predicted merged molecule derived from fragments bound at site 2.

2.5 Conclusion

In summary, this crystallographic fragment screen of IRF3CTD identifies 25 chemically diverse fragment hits at eight binding sites and provides, to our knowledge, the first set of experimentally validated small-molecule binders for IRF3. Three of these sites, the buried hydrophobic pocket at site 1, the C-terminal-tail-adjacent surface at site 2, and the adaptor- and dimer-binding hot spot at site 7, are particularly well placed to enable the development of chemical probes of IRF3 activation, of the host–pathogen interactions mediated by IRF3CTD, and ultimately of therapeutic candidates for the diseases in which aberrant IRF3 activity has been implicated.

2.6 Methods

Expression and purification of IRF3CTD. The IRF3CTD construct (residues 189–423) with a 6xHis tag followed by a TEV cleavage site at the N-terminus was purchased from Genscript in a pET-28a(+) vector. The protein was expressed in BL21(DE3) *E. coli* in LB medium at 30 °C overnight. Cells were harvested by centrifugation at 4,500 × *g* and resuspended in 20 mL of lysis buffer (50 mM Tris pH 8.0, 200 mM NaCl, 20 mM imidazole, Roche cOmplete Mini protease inhibitor) per liter of culture. Cells were lysed by sonication and clarified by centrifugation at 22,000 × *g*. The supernatant was filtered through a 0.45 μm membrane, loaded onto a 5 mL HisTrap FF column (Cytiva), washed with lysis buffer, and eluted with elution buffer (50 mM Tris pH 8.0, 200 mM NaCl, 500 mM imidazole). The eluate was digested with TEV protease overnight at 4 °C while dialyzing into TEV buffer (20 mM Tris pH 8.0, 100 mM NaCl, 20 mM imidazole, 5 mM glutathione). Glutathione was used in place of the more common DTT or TCEP because the latter reduce disulfide bonds essential to IRF3CTD stability, whereas glutathione does not. The digested protein was passed back over a HisTrap FF column to remove cleaved tags, uncleaved protein, and the His-tagged TEV. The flow-through was collected, concentrated, and filtered through a 0.22 μm PVDF membrane. Final polishing was performed on a Superdex 75 pg column (Cytiva) in buffer containing 20 mM Tris pH 8.0 and 100 mM NaCl.

Crystallization and fragment soaking. Purified IRF3CTD was concentrated to 6.3 mg/mL, with concentration determined by Nanodrop measurement of 10×-diluted samples back-calculated to the working concentration. Sitting-drop vapor-diffusion drops were set up in 3-well Midi plates (Hampton Research) using a Mosquito liquid handler (SPT Labtech). Each drop consisted of 200 nL of protein solution, 180 nL of crystallization solution (70 mM sodium acetate, 280 mM magnesium formate, pH 5.1), and 20 nL of crystal seeds in crystallization solution. The seed stock was prepared by crushing crystals from 15 drops containing many small crystals, adding them to crystallization solution along with 20% ethylene glycol, aliquoting, flash freezing in liquid nitrogen, and storing at −80 °C. Plates were incubated at 4 °C, and large crystals grew over the course of three weeks. Fragments

were soaked into the crystals using an Echo acoustic liquid handler (Beckman) by adding 90 nL of fragment stock and 500 nL of cryoprotectant solution (72% ethylene glycol, 140 mM sodium acetate, and 560 mM magnesium formate pH 5.1) directly to the crystal drop, yielding a final ethylene glycol concentration of 36% and a fragment concentration of 9 mM. Soaked crystals were incubated at 20 °C for two hours before being harvested and flash-cooled in liquid nitrogen.

Data collection and fragment analysis. Diffraction data were collected at Advanced Light Source beamline 8.3.1 and Stanford Synchrotron Radiation Lightsource beamline 12-1 at 100 K. Data reduction was performed with the DIALS pipeline [64] in xia2 [65]. Because the soaked data sets are essentially isomorphous, initial phase estimates were obtained directly from a common reference model (a DMSO-only soaked IRF3CTD crystal) rather than by molecular replacement. The reference coordinates were placed into the corresponding unit cell for each data set, and a single R_{free} flag set was propagated across the series so that the same test reflections were excluded throughout. Each data set was then refined in Phenix [66, 67], using a rigid-body refinement stage followed by restrained refinement of individual coordinates and atomic displacement parameters. Data sets with R_{free} below 0.35 were analyzed for fragment hits with PanDDA [38]. An initial analysis used a random subset of 34 high-resolution data sets to estimate the ground-state electron density corresponding to the ligand-free protein, which was subsequently used to identify crystals with fragment hits. Iterative analysis, excluding data sets that were identified as hits from ground-state characterization, produced a final set of 28 candidate hits. Fragments were placed into PanDDA event maps in Coot [68] using `pandda.inspect`. After fragment placement, the resulting structures were refined in Phenix following the protocol described in the PanDDA documentation, where ground state and bound state models are superimposed in a single PDB. The event maps of three data sets were too ambiguous to model, resulting in a final set of 25 fragment-bound structures. These fragment models were validated using metrics described by Pearce et al. (2017) [38]. Briefly, a Pearson's correlation coefficient, referred to as the real-space correlation coefficient (RSCC), reports on the similarity between the calculated electron density from the fragment

model and the observed electron density from the data. Following Pearce et al.'s recommendation, we used a RSCC threshold of 0.7 as a minimum requirement for accepting a fragment model. A similar metric, the real-space Z-difference score, indicates the amount of electron density calculated from the fragment that does not align with the observed electron density, and vice versa, relative to the noise of the difference density map. We accepted values for this score of 3.0 or less, again aligning with Pearce et al. Both metrics were calculated using the CCP4 program EDSTATS [69].

Fragmenstein [63] was used to predict fragments that can be merged or linked to improve binding affinity. Fragmenstein analyzes the fragments for a given site together in the context of the binding pocket. It uses Rosetta energy scores to identify which combinations of linking, merging, and elaborating the fragments results in a lower $\Delta G_{\text{binding}}$ compared to the parent fragments.

Per-chapter acknowledgment

Chapter 2, in part, contains material currently being prepared for submission for publication as: **Rodriguez, J. M.**; Hernandez, V.; Barron, O.; Sweeney, M.; Tong, A.; Correy, G.; Thompson, M. C. "Crystallographic fragment screen of the IRF3 regulatory domain."

Chapter 3

Biophysical characterization of the proposed RSV NS1–IRF3 interaction

3.1 Introduction

The innate immune system is a critical, early-stage line of defense against invading viruses and includes an integrated set of cellular and molecular mechanisms that detect pathogen-associated molecular patterns, restrict replication, and clear infected cells before adaptive immunity engages [17]. Central to this antiviral response is Interferon Regulatory Factor 3 (IRF3), a transcription factor constitutively expressed in most human cells and held inactive in the cytoplasm under resting conditions. When viral nucleic acids are detected by pattern recognition receptors, such as TLR3 or RIG-I, signaling cascades converge on IRF3 and trigger its activation, leading to dimerization, translocation to the nucleus, recruitment of coactivators, and binding to interferon promoters. The resulting transcription of type I interferons coordinates a broad antiviral state across infected and neighboring cells [70].

A wide range of viruses encode proteins that interfere with IRF3 to evade the innate immune response. Documented examples span viral families, includ-

ing SARS-CoV-2 [61], respiratory syncytial virus [32], and influenza A [57], with strategies ranging from blocking phosphorylation to redirecting IRF3 for proteasomal or lysosomal degradation [54, 55, 33, 56, 25, 58, 59, 60, 24]. Despite this breadth of cellular evidence, the binding interactions between viral proteins and IRF3 remain poorly characterized at the biophysical level. The detection methods of the interactions listed above rely largely on co-immunoprecipitation from mammalian cells followed by Western blots. This qualitative approach cannot rule out the possibility that other interacting partners are required for the association between the viral protein and IRF3. Quantitative binding measurements with purified components are needed to confirm a direct interaction.

Here, we sought to quantify the binding of RSV NS1 to IRF3. NS1 was selected as a good candidate for these experiments based on prior work showing that NS1 blocks IRF3 activation in RSV-infected cells [32]. Additionally, the crystal structure of NS1 has been determined, providing a starting point for biophysical characterization [71]. We used the C-terminal domain (CTD) of IRF3 as our study target based on immunoprecipitation experiments that showed that RSV NS1 associates with both IRF3 and the coactivator CBP in infected cells [33]. The IRF3-CBP interaction occurs at the CTD [49], so if NS1 disrupts IRF3-CBP association, the CTD is the most likely site of NS1 engagement. Additionally, the CTD is the site of many regulatory events in IRF3 activation, including phosphorylation by TBK1 and IKKε, dimerization, and recruitment of CBP and p300. This makes the CTD an attractive target for viral interference. For example, rotavirus NSP1 binds at the site on IRF3CTD where multiple host adaptor proteins bind [23]. Whether RSV NS1 engages the CTD directly, however, has not been tested with purified components. To address this gap, we tested for direct NS1-IRF3CTD binding using analytical size-exclusion chromatography, fluorescence polarization, and biolayer interferometry.

3.2 Results

3.2.1 Analytical Size-Exclusion Chromatography

As an initial test for a direct interaction between the phosphorylated IRF3CTD (pIRF3CTD) and NS1, the proteins were analyzed by SEC. IRF3CTD was phosphorylated at Ser386 by genetic code expansion [72], using an amber stop codon in place of the Ser386 codon to direct site-specific phosphoserine incorporation. Phosphorylation was confirmed with mass spectrometry. Equivalent molar quantities of pIRF3CTD and unphosphorylated IRF3CTD were resolved on a size-exclusion column. The two samples eluted at distinct volumes consistent with a dimeric pIRF3CTD species and a monomeric IRF3CTD species (Fig. 3.1). When NS1 was combined with pIRF3CTD at a 1:1 molar ratio and applied to the same column, the elution profile changed significantly. Rather than eluting at the dimer volume, the pIRF3CTD + NS1 sample eluted near the monomer volume, but with a small shift to a lower elution volume. If NS1 had bound the intact pIRF3CTD dimer without disrupting it, we would expect only a modest shift of the dimer peak. Instead, we observed that the dimer peak was absent and the mixed sample eluted near the monomer position. This could indicate that NS1 binds pIRF3CTD in a manner that disrupts the dimer interface, freeing pIRF3CTD monomers. NS1 then remains partially bound to the freed monomers, producing the small additional shift to lower elution volume relative to bare IRF3CTD. This result is consistent with a direct interaction and motivated quantitative binding measurements.

3.2.2 Fluorescence Polarization

To quantify the apparent interaction between NS1 and IRF3CTD, we turned to fluorescence polarization (FP). We labeled NS1 with BODIPY using maleimide-reactive chemistry and titrated IRF3CTD or pIRF3CTD into a fixed concentration of BODIPY-NS1. The monomeric IRF3CTD bound BODIPY-NS1 with an apparent K_{D} of 0.151 μM (Fig. 3.2A), and the dimeric pIRF3CTD bound with an apparent K_{D} of 0.381 μM (Fig. 3.2B). The monomeric IRF3CTD data fit best to a Hill coefficient of 1, while the pIRF3CTD data fit best to a Hill coefficient of 2.

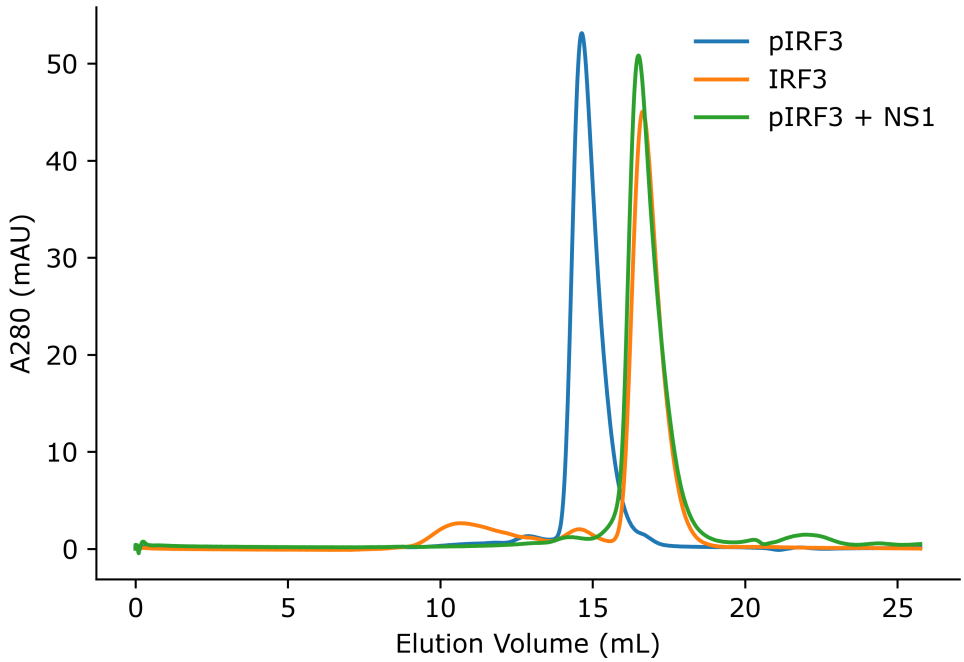


Figure 3.1: **NS1 disrupts the pIRF3CTD dimer in size-exclusion chromatography.** pIRF3CTD, IRF3CTD, and a 1:1 molar combination of pIRF3CTD and NS1 were resolved on a Superdex 75 column. pIRF3CTD alone elutes at the dimer position (blue line), while IRF3CTD alone elutes at the monomer position (orange line). The pIRF3CTD + NS1 sample elutes near the monomer position with a small additional shift to a lower elution volume, consistent with NS1 disrupting the dimer and remaining bound to the freed monomers (green line).

The cooperativity observed with the dimer is consistent with the SEC result, in which NS1 appears to disrupt the pIRF3CTD dimer. Binding of one NS1 molecule to a dimer subunit may destabilize the dimer interface, freeing the second subunit and exposing a second binding site for cooperative engagement of a second NS1 molecule.

To test whether the apparent binding was specific, we performed a negative control assay using lysozyme in place of IRF3CTD. Lysozyme has no known biological connection to NS1 and is routinely used as an inert control protein. Unexpectedly, NS1 bound lysozyme with an apparent K_{D} of 0.393 μM (Fig. 3.2C), similar to the IRF3CTD measurements. Fluorophores have a known propensity for nonspe-

cific surface interactions, so we initially considered the possibility that this result reflected IRF3 interacting with the fluorophore rather than NS1. This concern motivated our next experiment, which used biolayer interferometry (BLI) to test for binding without a fluorescent label.

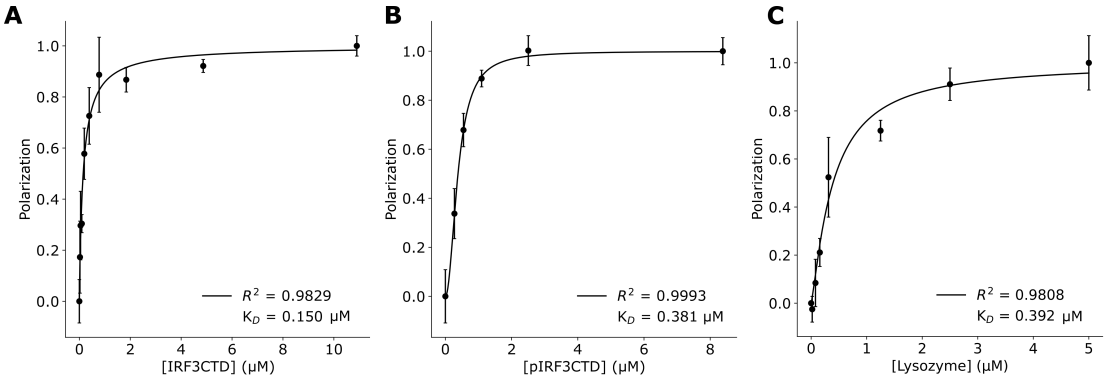


Figure 3.2: **Fluorescence polarization shows that NS1 binds IRF3CTD with similar affinity to lysozyme.** **(A)** BODIPY-labeled NS1 binds the monomeric IRF3CTD with an apparent K_{D} of 0.151 μM ($\pm$0.016), fit with a Hill coefficient of 1. **(B)** BODIPY-labeled NS1 binds the dimeric pIRF3CTD with an apparent K_{D} of 0.381 μM ($\pm$0.00717), fit with a Hill coefficient of 2. **(C)** 5FAM-labeled NS1 binds lysozyme in the negative control assay with an apparent K_{D} of 0.393 μM ($\pm$0.0538).

3.2.3 Biolayer Interferometry

To test the NS1-IRF3 interaction without a fluorescent label, we used BLI. Biotin was attached to either IRF3CTD or NS1 using sortase-mediated labeling, allowing the labeled protein to be immobilized on a streptavidin probe while the other binding partner remained free in solution. Two assays were performed. The first tested NS1 binding to immobilized IRF3CTD. The second tested binding of phosphorylated, full-length IRF3 (pIRF3FL) to immobilized NS1. pIRF3FL was used here because it is representative of the activated IRF3 species observed in prior cellular studies. In the first assay with biotin-IRF3CTD, binding was detected with an apparent K_{D} of 6.73 nM, which indicates a strong interaction (Fig. 3.3, blue). However, the negative control (no biotin-IRF3CTD loaded) produced binding behavior nearly indistinguishable from the IRF3CTD-loaded probe, with an apparent

K_D of 12.6 nM (Fig. 3.3, gray). This result suggests that NS1 binds nonspecifically to the streptavidin probe surface, and that the apparent binding to IRF3CTD is dominated by this nonspecific association. In the second assay, using immobilized NS1, no binding to pIRF3FL was detected (Fig. 3.4). This rules out potential interference of the biotin-IRF3 orientation to the probe and the possibility that NS1 only binds the non-phosphorylated IRF3. Taken together, the nonspecific signal in the first assay and the absence of any signal in the second assay indicate that there is no specific NS1-IRF3 interaction detectable *in vitro*.

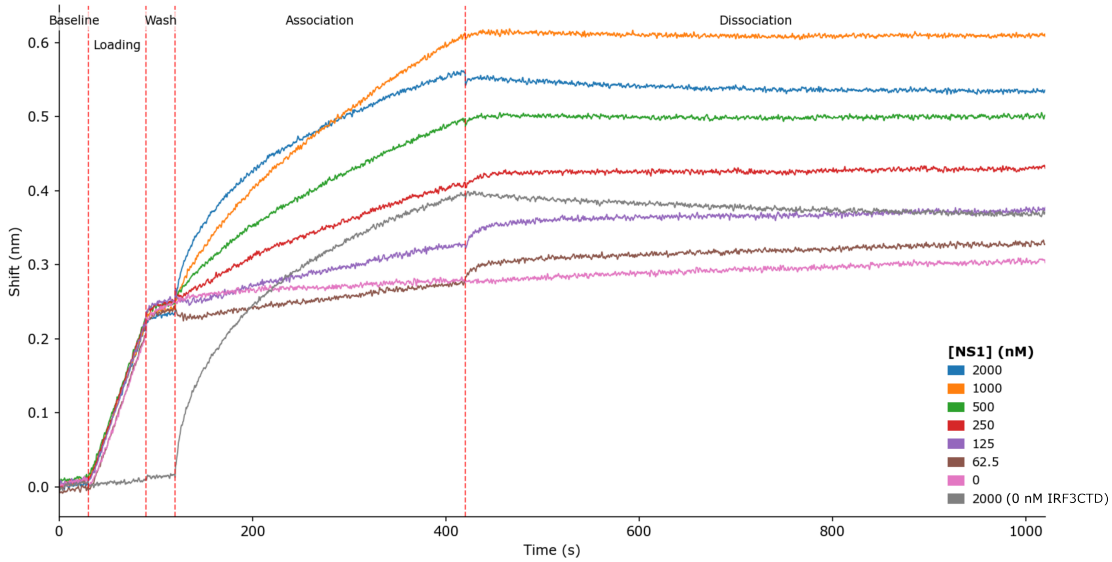


Figure 3.3: **BLI with biotin-IRF3CTD and NS1 reveals nonspecific binding of NS1 to the probe surface.** Biotin-labeled IRF3CTD at 50 nM was loaded onto a streptavidin BLI probe and dipped into NS1 solutions ranging from 62.5 nM to 2 μM (blue). The negative control was a streptavidin probe with no biotin-IRF3CTD loaded, dipped into 2 μM NS1 (gray). Each trace includes baseline, loading, association, and dissociation phases. The negative control produces association and dissociation behavior nearly identical to the IRF3CTD-loaded probe, indicating that the binding signal is dominated by nonspecific NS1 association with the probe surface.

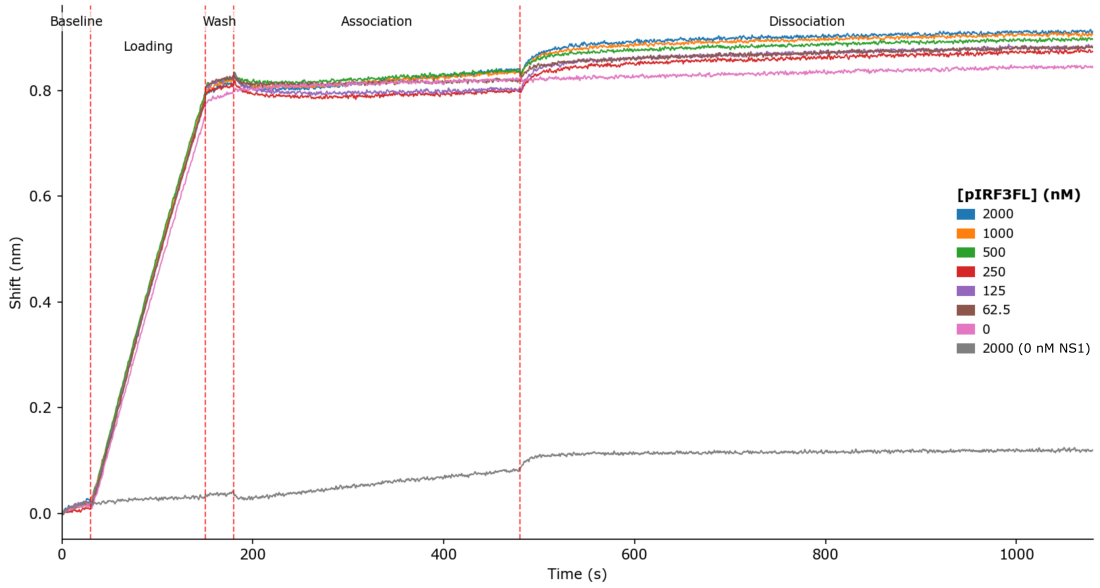


Figure 3.4: **No binding is detected between biotin-NS1 and pIRF3FL.** Biotin-labeled NS1 was loaded onto a streptavidin BLI probe and dipped into phosphorylated, full-length IRF3 (pIRF3FL) solutions ranging from 62.5 nM to 2 μM. No binding was observed at any concentration tested.

3.3 Discussion

This study set out to determine whether RSV NS1 binds the IRF3 C-terminal domain directly, with the goal of characterizing IRF3 as a potential broad-spectrum antiviral target. The literature on viral interference with IRF3 is substantial in cellular and immunological terms, but biophysical and structural characterization of direct viral protein-IRF3 interactions is limited to two cases. Rotavirus NSP1 binds the IRF3 CTD through a pLxIS motif at the same site used by the host adaptor proteins STING, MAVS, and TRIF, with SPR affinities of $\sim$16 μM for phosphorylated NSP1 and $\sim$200 μM for the unphosphorylated form [23]. The HPV-16 E6 oncoprotein engages a different surface, binding an LxxLL motif in the N-terminal DNA-binding domain of IRF3 [24]. RSV NS1 is a particularly attractive target for biophysical characterization among the viral proteins implicated in IRF3 inhibition, for three reasons. First, RSV is a major pediatric pathogen with no available antiviral therapeutics, so disrupting NS1-mediated suppression of IRF3 could restore innate immune signaling during infection. Second, while Spann

et al. established the cellular phenotype that NS1 and NS2 cooperatively block IRF3 activation in RSV-infected cells [32], Ren et al. proposed that NS1 disrupts the IRF3-CBP interaction [33], a mechanism that is directly testable with purified components. Third, NS1 has a published crystal structure [71], establishing that the protein is amenable to recombinant expression and purification. Despite these favorable starting conditions, no biophysical or structural characterization of the NS1-IRF3 interaction has been published. Our work aimed to validate and quantify this interaction.

Our results from analytical SEC were consistent with a direct interaction. NS1 disrupted the pIRF3CTD dimer and shifted the elution position of the resulting monomers, behavior expected if NS1 binds the dimerization interface or otherwise destabilizes the dimer. Encouraged by this result, we pursued FP and BLI to quantify binding. In FP, NS1 bound both monomeric IRF3CTD and dimeric pIRF3CTD with apparent submicromolar affinities, but it also bound the negative control protein lysozyme with nearly identical affinity. In the first BLI assay, NS1 produced an apparent nanomolar interaction with immobilized IRF3CTD, but it bound the probe surface itself with roughly the same affinity in the absence of IRF3CTD. The second BLI assay failed to detect an interaction between the physiologically relevant pIRF3FL and NS1. Taken together, these observations indicate that the apparent binding signals reflect nonspecific behavior of NS1 rather than a specific biological interaction with IRF3.

Two alternative explanations are worth considering. NS1's propensity for nonspecific binding, demonstrated by its binding to lysozyme and to the BLI probe surface, could have been responsible for the positive signal in previously reported co-immunoprecipitation experiments. NS1 could have associated nonspecifically with IRF3 directly, or with any of the many proteins IRF3 complexes with during activation, including its coactivator CBP. The co-immunoprecipitation followed by western blot experiments performed previously cannot distinguish a specific direct interaction from indirect or nonspecific association under those conditions. A separate concern, that NS1 binds the N-terminal DNA-binding domain rather than the CTD, does not account for our observations. If NS1 had bound the

N-terminal domain, the second BLI assay with pIRF3FL (which contains both domains) would have shown binding. It did not. Combined with the lysozyme and probe-surface controls, this argues that NS1 does not have a specific binding interface for any epitope of IRF3 that can be detected by these methods.

The path forward for characterizing specific viral protein-IRF3 interactions should begin with methods to identify these interactions using purified components, such as *in vitro* pull-down assays. Confirming that a viral protein and IRF3 interact outside the cellular context is a necessary prerequisite to quantitative biophysical characterization. Only after such an interaction is established does it become reasonable to invest in solution binding assays for K_{D} determination or in structural studies of a complex. Our results do not rule out a specific NS1-IRF3 interaction in cells, but they do indicate that any such interaction either requires additional cellular factors or occurs under conditions we could not reproduce *in vitro*. While the immediate finding here is negative for specific NS1-IRF3 binding, this work demonstrates the importance of orthogonal controls when characterizing viral protein interactions with IRF3, and points toward *in vitro* pull-down assays with purified components as a necessary next step before downstream biophysical and structural studies.

3.4 Methods

Protein Expression and Purification. IRF3CTD was expressed and purified as described in Chapter 2. Phosphorylation of IRF3CTD was achieved with a genetic code expansion system provided by GCE4All (Oregon State University) [72]. Briefly, the codon for Ser386 was mutated to the amber stop codon (TAG) using PCR-based mutagenesis. The IRF3CTD construct was then cloned into the pRBC expression plasmid provided by GCE4All and verified by full-plasmid sequencing. The plasmid was co-transformed into an RF1-deficient B95 *E. coli* strain along with the pKW2-EFSep plasmid, which encodes the phosphoserine incorporation machinery. Expression and purification of pIRF3CTD followed the same protocol as IRF3CTD, aside from the culture media composition, which contained

ZY, 2 mM magnesium sulfate, 0.5% α-D-glucose, 25 mM sodium dibasic phosphate, 25 mM potassium monobasic phosphate, 50 mM ammonium chloride, and 5 mM sodium sulfate. pIRF3FL was cloned, expressed, and purified following the same protocol as pIRF3CTD.

The NS1 construct was purchased from GenScript in a pET-22b(+) vector. The plasmid was transformed into BL21 (DE3) *E. coli* and expressed using 0.5 mM IPTG in TB media at 18 °C overnight. The cells were harvested by centrifuging at 4,500 $\times$ g and lysed via sonication in lysis buffer (20 mM Tris pH 8, 300 mM sodium chloride, 20 mM imidazole, 2 mM TCEP, 5% (v/v) glycerol, and mini cOmplete protease inhibitor cocktail). The lysate was clarified by centrifuging at 22,000 $\times$ g and the supernatant was filtered with a 0.45 μm PVDF filter. NS1 was then purified using a HisTrap FF column (Cytiva), eluting with elution buffer (20 mM Tris pH 8, 300 mM sodium chloride, 500 mM imidazole, 2 mM TCEP, 5% (v/v) glycerol). The eluted NS1 sample was then digested overnight at 4 °C with TEV in TEV buffer (20 mM Tris pH 7.5, 150 mM sodium chloride, 5% (v/v) glycerol, and 1 mM DTT). The digested sample was applied to a HisTrap FF column to remove cleaved His-tag and His-tagged TEV and buffer exchanged into a final buffer containing 20 mM Tris pH 7.5, 150 mM NaCl, and 5% glycerol. The protein was concentrated and polished with a Superdex 75 pg 16/600 column (Cytiva).

Analytical Size Exclusion Chromatography. Equivalent molar quantities of IRF3CTD, pIRF3CTD, and a 1:1 molar combination of pIRF3CTD and NS1 were resolved on a Superdex 75 Increase 10/300 GL column (Cytiva). The column was run at 0.75 mL/min at 4 °C.

Fluorophore and Biotin Labeling. NS1 was labeled with the fluorophores BODIPY and 5-carboxyfluorescein containing maleimide groups. NS1 was concentrated to 5 μM (measured by BCA assay against a BSA standard from a Thermo Scientific BCA kit) and 900 μL was combined with 100 μL of a reaction solution containing 500 μM fluorophore (0.22 μm filtered, concentration determined by NanoDrop absorbance), 20 mM Tris pH 7.5, 150 mM sodium chloride, 5 mM EDTA, and 2 mM TCEP. The labeling solution was incubated at 4 °C overnight,

then combined with dye-removal resin and spun through a filter. The flow through was then run over a Superdex 75 Increase 10/300 GL column (Cytiva). Labeling was confirmed using MALDI-MS.

For biotin labeling, peptides containing the sortase recognition motif attached to biotin (biotin-LPETGG) were purchased from GenScript. NS1 and IRF3CTD were labeled following the protocol described by Theile et al. [73]. Briefly, 500 μL of 100 μM NS1 or IRF3CTD was combined with 500 μL of a reaction buffer containing NS1 SEC buffer plus 20 mM calcium chloride, 1 mM peptide, and 250 μM sortase A. The solution was incubated at 4 °C for 24 hr. The sample was then passed through a HisTrap FF column to remove sortase and polished on a Superdex 75 Increase 10/300 GL column to remove excess peptide.

Fluorescence Polarization Assay. FP binding measurements rely on a large change in rotational correlation time upon binding. NS1 (17 kDa) is sufficiently small compared to IRF3CTD (26 kDa), and the excitation lifetimes of BODIPY and 5FAM are sufficiently long, to exhibit a large change in polarization upon binding. IRF3CTD titrations were made by serial twofold dilutions from a 5 μM stock (measured by NanoDrop A_{280}), ending at 40 nM after seven dilutions. The labeled NS1 sample was concentrated to 250 nM, determined by fluorescence intensity due to the low A_{280} and BCA signals at that concentration. 25 μL of NS1 sample was added to 225 μL of each IRF3CTD titration and mixed by pipetting. 30 μL of the mixed solution was pipetted into each of seven wells in a row on a 384-well black opaque NBS plate (Corning cat. no. 3820). Controls were made in a similar manner, including (−)IRF3CTD, (−)NS1, and buffer only. Total, perpendicular, and parallel fluorescence intensities were measured on a CLARIOstar plate reader (BMG LabTech). The buffer blank was subtracted from all samples. Because IRF3CTD exhibits intrinsic fluorescence, perpendicular and parallel intensities from the (−)NS1 measurements were subtracted from the IRF3CTD + NS1 measurements. Outliers in total, perpendicular, or parallel intensity were identified and rejected using an inter-quartile range cutoff with a threshold of 2.22, corresponding to approximately 3 standard deviations. The filtered data were fit to a standard Hill equation with a Hill coefficient of 1 or 2, depending on which

yielded the lower residual error.

Biolayer Interferometry Assay. The biotin-IRF3CTD sample for BLI was made by diluting a 50 μM stock to 50 nM in Q buffer (GatorBio). NS1 titrations were made by serially diluting NS1 stock from 2 μM to 62.5 nM in Q buffer. The same concentrations for sample loading (50 nM) and titrations (2 μM to 62.5 nM) were used for the assay with pIRF3FL, where biotin-NS1 was immobilized on the probe. BLI measurements were collected on a Gator Prime system (GatorBio) using streptavidin (SA) probes (GatorBio). Negative controls consisted of probes with no biotin-protein loaded, dipped into the maximum analyte concentration. Binding analysis was performed with GatorBio software.

Per-chapter acknowledgment

Chapter 3 was authored by **Rodriguez, J. M.**, with contributions from Barron, O. G.; Meyer, K.; Thompson, M. C.

Chapter 4

Crystallography-informed computational protein design

This chapter presents two collaborative projects with the Chica group at the University of Ottawa, in which crystallographic information about protein conformational heterogeneity is used to inform and validate the computational design of enzymes. Part A describes the multistate-design remodeling of the conformational landscape of *E. coli* aspartate aminotransferase to alter substrate selectivity. Part B describes ensemble-based design of two *de novo* Kemp eliminases to enhance catalytic efficiency.

4.1 Part A: Computational remodeling of an enzyme conformational landscape for altered substrate selectivity

Abstract

Structural plasticity of enzymes dictates their function. Yet, our ability to rationally remodel enzyme conformational landscapes to tailor catalytic properties remains limited. Here, we report a computational procedure for tuning conformational landscapes that is based on multistate design of hinge-mediated domain

motions. Using this method, we redesign the conformational landscape of a natural aminotransferase to preferentially stabilize a less populated but reactive conformation and thereby increase catalytic efficiency with a non-native substrate, resulting in altered substrate selectivity. Steady-state kinetics of designed variants reveals activity increases with the non-native substrate of approximately 100-fold and selectivity switches of up to 1900-fold. Structural analyses by room-temperature X-ray crystallography and multitemperature nuclear magnetic resonance spectroscopy confirm that conformational equilibria favor the target conformation. Our computational approach opens the door to targeted alterations of conformational states and equilibria, which should facilitate the design of biocatalysts with customized activity and selectivity.

4.1.1 Introduction

Enzymes are flexible macromolecules that sample multiple structural states, described by a conformational energy landscape [74]. It is the relative stability of these conformational states, and the ability of enzymes to transition between them, that ultimately dictates enzymatic function [75, 76, 77]. Analyses of directed evolution trajectories have shown that evolution can reshape enzyme conformational landscapes by enriching catalytically productive states and depopulating non-productive ones [78, 79, 80], leading to enhanced catalytic activity. Similar mechanisms contribute to the evolution of substrate selectivity, where transient conformations responsible for activity on non-cognate substrates become enriched [81, 82, 83]. Thus, it should be possible to harness the pre-existing conformational plasticity of enzymes to tailor their catalytic properties by reshaping their conformational landscapes. However, predicting the effect of mutations on these energy landscapes remains challenging, and current computational enzyme design protocols, which focus on a single structural state, are poorly optimized for this task. New design methodologies are therefore required for the targeted alteration of subtle conformational states and equilibria, which would in turn facilitate the design of biocatalysts with customized activity and selectivity.

Here, we report a computational procedure for rationally tuning enzyme con-

formational landscapes that is based on multistate computational protein design, a methodology that allows protein sequences to be optimized on multiple structural states [45]. As a case study, we remodel the conformational landscape of aspartate aminotransferase, an enzyme that switches between open and closed conformations via hinge movement, which involves the rotation of a protein domain relative to another around an axis between two planes. Using our approach, we enrich the less populated but catalytically active closed conformation in order to increase catalytic efficiency ($k_{\mathrm{cat}}/K_{\mathrm{M}}$) with the non-native substrate L-phenylalanine, leading to altered substrate selectivity. Steady-state kinetics reveal $k_{\mathrm{cat}}/K_{\mathrm{M}}$ increases of up to 100-fold towards this aromatic amino acid, resulting in a selectivity switch of up to 1900-fold, and structural analyses by room-temperature X-ray crystallography and multitemperature nuclear magnetic resonance (NMR) spectroscopy confirm that the conformational landscape is remodeled to favor the target state. Our methodology for altering conformational equilibria should be applicable to many enzymes and proteins that undergo hinge-mediated domain motions.

4.1.2 Results

Computational remodeling of conformational landscape

E. coli aspartate aminotransferase (AAT) is a pyridoxal phosphate (PLP)-dependent enzyme that catalyses the reversible transamination of L-aspartate with α-ketoglutarate, yielding oxaloacetate and L-glutamate. During its catalytic cycle, AAT undergoes hinge movement to switch between an open conformation, in the ligand-free form, and a closed conformation upon association with substrates or inhibitors (Fig. 4.1a,b) [84]. Previously, AAT was redesigned to change its substrate specificity to allow transamination of aromatic amino acids [46]. To do so, six of the 19 residues that are strictly conserved in AAT enzymes were replaced by those found at corresponding positions in the homologous *E. coli* tyrosine aminotransferase. These residues were selected because they line the substrate channel leading to the catalytic pocket or are located near the cofactor phosphate moiety. This process yielded a hexamutant (HEX) of AAT that is approximately two orders of magnitude more catalytically efficient than the wild type (WT) with the aromatic

amino acid L-phenylalanine (Table 4.1), and similarly efficient for transamination of L-aspartate. Unexpectedly, it was found that unlike WT, HEX was closed in its ligand-free form (Fig. 4.1c) [47], demonstrating that those six mutations shifted its conformational equilibrium to favor the closed state (Fig. 4.1d), which is the active conformation of the enzyme [85]. This equilibrium shift was accompanied by enhanced L-phenylalanine transamination activity resulting from a larger increase in affinity for this non-native substrate than for the native L-aspartate substrate [46]. Since domain closure in AAT is driven by the excess binding energy of the native substrate L-aspartate, which forms strong polar interactions with two active-site arginines [85, 86], the observed equilibrium shift towards the closed state in the absence of ligand was proposed to facilitate access to this active conformation for non-native substrates whose binding may not provide sufficient energy to induce domain closure due to their lower structural complementarity to the active site [47]. Based on these observations, we postulated that we could rationally remodel the conformational landscape of AAT by using multistate computational protein design [45] to identify novel mutation combinations that can preferentially stabilize the closed conformation over the open conformation, and in doing so, increase catalytic efficiency and selectivity for L-phenylalanine.

To test this hypothesis, we implemented a computational strategy (Fig. 4.2) that proceeds in five steps: (1) identification of hinge-bending residues involved in transition between open and closed conformations; (2) generation of structural ensembles approximating backbone flexibility to model open and closed conformational states; (3) optimization of side-chain rotamers for all allowed amino-acid combinations at key hinge-bending residues and neighboring positions, on each ensemble; (4) calculation of energy differences between open and closed states to predict preferred conformation, and (5) combinatorial library design using computed energy differences to select mutant sequences for experimental testing.

To identify hinge-bending residues for the open/closed conformational transition, crystal structures of WT AAT in its open and closed forms (PDB ID: 1ARS and 1ART, respectively [84]) were used as input for hinge movement analysis with DynDom, a program that identifies domains, hinge axes, and hinge bend-

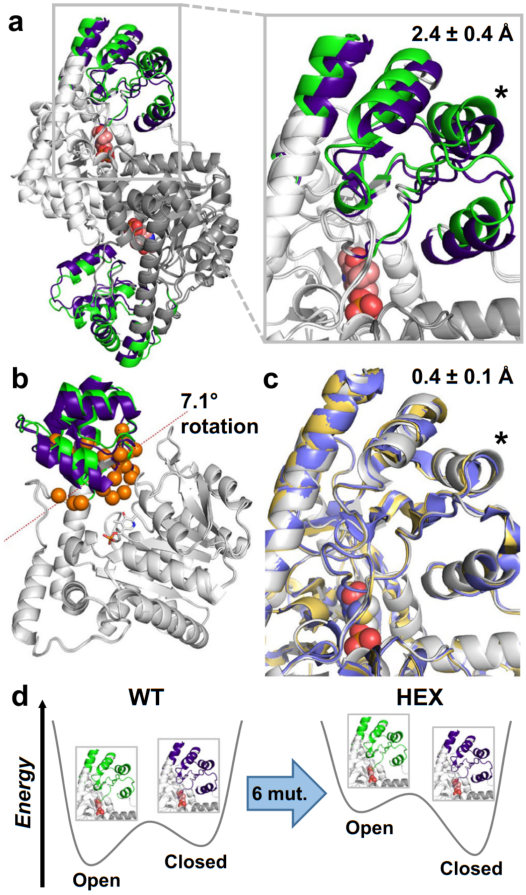


Figure 4.1: **AAT conformational landscape.** **(a)** *E. coli* AAT is a 90 kDa homodimer that undergoes a conformational change from an open (green, PDB ID: 1ARS) to closed (dark blue, PDB ID: 1ART) state upon substrate binding. This conformational transition involves rotation of a small moving domain (colored) relative to a fixed domain (white and gray for chains A and B, respectively), which causes a 2.4 ± 0.4 Å displacement (mean Cα distance ± s.d.) of the helix formed by residues K355–F365 (indicated by an asterisk). The PLP cofactor bound at the active site is shown as spheres (salmon), and is present in both conformations. **(b)** Hinge movement analysis of chain A reveals a 7.1-degree rotation of the moving domain relative to the fixed domain along an axis between two planes (dotted line). Hinge-bending residues and PLP are shown as orange spheres and white sticks, respectively. **(c)** Superposition of HEX structures in the absence (yellow, PDB ID: 1AHE) and presence (blue, PDB ID: 1AHY) of bound inhibitor (representation not shown) with that of the WT closed state (white, PDB ID: 1ART) show that this mutant is closed in both cases. Cα displacement (mean ± s.d.) of residues K355–F365 (asterisk) is indicated. **(d)** These results demonstrate that the six mutations (mut.) of HEX remodel its conformational landscape to favor the closed conformation.

ing residues in proteins for which two conformations are available [87]. DynDom analysis revealed a small moving domain (Fig. 4.1b) that rotates by 7.1 degrees about the hinge axis from the larger fixed domain, and identified 25 hinge-bending residues. We selected two of these residues for design, Val35 and Lys37 (numbering based on Uniprot sequence P00509), because they are found on the flexible loop connecting the moving and fixed domains (Fig. 4.2). We also selected for design residues Thr43 and Asn64, which are not part of the hinge, but whose side chains form tight packing interactions with those of Val35 and Lys37. Interestingly, these residues comprise four of the six positions that were mutated in HEX (Table 4.1), demonstrating that our analysis using only WT structures led to the identification of positions that contribute to controlling the open/closed conformational equilibrium in AAT.

Next, we generated backbone ensembles from the open- and closed-state crystal structures to approximate the intrinsic flexibility of these two conformational states using the PertMin algorithm [88], which we previously showed to result in improved accuracy of protein stability predictions when used as templates in multistate design [89]. Using the protein design software Phoenix [90, 91], we optimized rotamers for all combinations of proteinogenic amino acids with the exception of proline at the four designed positions on each backbone ensemble, yielding Boltzmann-weighted average energies for 130,321 (19^4) AAT sequences that reflect their predicted stability on each conformational state. To identify mutant sequences that preferentially stabilized the closed conformation, we computed the energy difference between closed and open state ensembles ($\Delta E = E_{\text{closed}} - E_{\text{open}}$) for each sequence using the Phoenix potential energy function developed for protein design. Phoenix energies correspond to the sum of pairwise nonbonded interaction energies between rotamers and between rotamers and template in the folded state, without consideration of bonded interactions, entropy or the energy of the unfolded state. As a final step, we used these ΔE values as input to the CLEARSS library design algorithm [90] to generate a 24-member combinatorial library of AAT mutants predicted to favor the closed state (Closed library) with a range of values (-86.0 to -9.5 kcal mol^{-1}) encompassing that of HEX (-76.8 kcal mol^{-1},

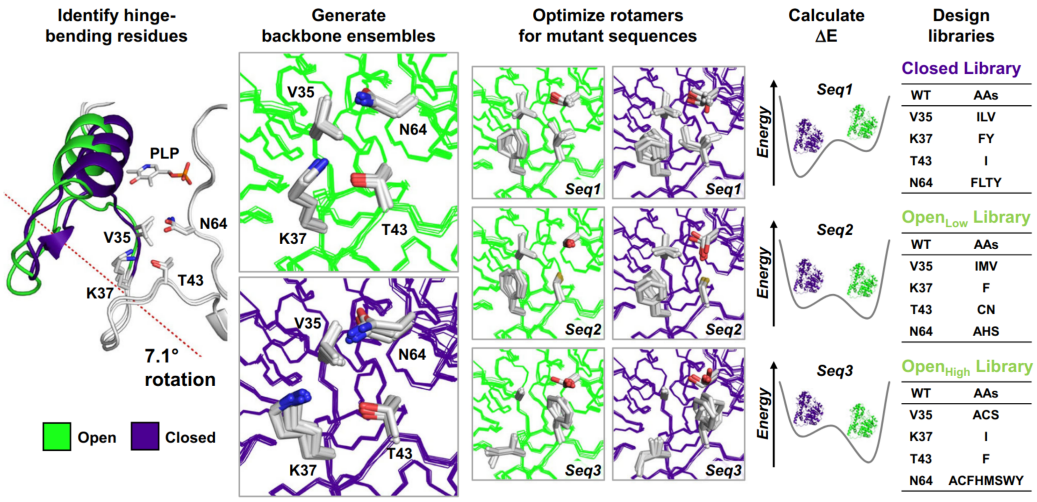


Figure 4.2: **Computational remodeling of AAT conformational landscape by multistate design.** To remodel the AAT conformational landscape, we followed a 5-step process: (1) identification of hinge-bending residues involved in transition between open (green) and closed (dark blue) conformational states; (2) generation of structural ensembles approximating backbone flexibility to model open and closed states; (3) optimization of rotamers for mutant sequences on both open- and closed-state ensembles; (4) calculation of energy differences between conformational states ($\Delta E = E_{\text{closed}} - E_{\text{open}}$) to predict equilibrium of each mutant, and (5) combinatorial library design using ΔE values to generate Closed, Open_{Low}, and $\text{Open}_{\text{High}}$ libraries for experimental testing. Designed residues V35, K37, T43, and N64 correspond to V39, K41, T47, and N69 in the previously published crystal structures of wild-type AAT (PDB ID: 1ARS and 1ART).

Table 4.2). As controls, we also generated two libraries of sequences predicted to favor the open conformation: the Open_{Low} library, which contains 18 sequences predicted to stabilize the open state with ΔE values (0.8–16.4 kcal mol^{-1}) comparable to that of the WT (14.2 kcal mol^{-1}, Table 4.2), and the $\text{Open}_{\text{High}}$ library, which contains 24 sequences predicted to more strongly favor the open state due to substantial destabilization of the closed state by >120 kcal mol^{-1}. While we postulated that the Open_{Low} library would yield mutants with wild-type-like conformational landscapes and therefore similar catalytic efficiency and substrate selectivity, we hypothesized that $\text{Open}_{\text{High}}$ library mutants would be less efficient than WT with both native and non-native substrates due to their strong destabilization of the closed conformation, which is the active form of the enzyme [85].

Thus, experimental characterization of these three mutant libraries, which comprise non-overlapping sequences (Fig. 4.2), allowed us to assess the ability of the ΔE metric to predict sequences with conformational landscapes favoring the open or closed states.

Table 4.1: Apparent kinetic parameters of *E. coli* AAT and its mutants for transamination of various amino-acid donors with α-ketoglutarate as acceptor.

		Mutations[b]				L-Aspartate[c]			
Enzyme[a]	Library	V35	K37	T43	N64	K_M (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (M^{-1} s^{-1})	Selectivity[d]
Controls									
WT	—	–	–	–	–	0.21 ± 0.03	7.8 ± 0.3	37,000 ± 5000	0.01
HEX[e]	—	L	Y	I	L	0.059 ± 0.007	1.32 ± 0.03	22,000 ± 3000	1.5
Closed Library									
IYIT	Closed	I	Y	I	T	0.014 ± 0.002	0.155 ± 0.005	11,000 ± 2000	2.7
VFIT	Closed	–	F	I	T	0.027 ± 0.003	0.90 ± 0.01	33,000 ± 4000	0.8
VFIY	Closed	–	F	I	Y	0.12 ± 0.01	0.263 ± 0.005	2200 ± 200	19
VYIT	Closed	–	Y	I	T	0.08 ± 0.01	0.37 ± 0.01	4600 ± 600	8
VYIY	Closed	–	Y	I	Y	0.09 ± 0.02	0.244 ± 0.006	2700 ± 600	13
$Open_{Low}$ Library									
IFCA	$Open_{Low}$	I	F	C	A	0.031 ± 0.004	2.63 ± 0.06	80,000 ± 10,000	0.25
MFCA	$Open_{Low}$	M	F	C	A	0.018 ± 0.003	1.21 ± 0.02	70,000 ± 10,000	0.47
VFCA	$Open_{Low}$	–	F	C	A	0.050 ± 0.006	4.4 ± 0.1	90,000 ± 10,000	0.12
VFCS	$Open_{Low}$	–	F	C	S	0.068 ± 0.007	6.0 ± 0.1	88,000 ± 9000	0.08

continued on next page

Table 4.1 (continued)

Enzyme[a]	Library	Mutations[b] V35	K37	T43	N64	L-Aspartate[c] K_M (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (M^{-1} s^{-1})	Selectivity[d]
$Open_{High}$ Library									
AIFS	$Open_{High}$	A	I	F	S	0.32 ± 0.03	2.42 ± 0.04	7600 ± 700	0.03
CIFC	$Open_{High}$	C	I	F	C	0.13 ± 0.01	2.62 ± 0.04	20,000 ± 2000	0.03
CIFS	$Open_{High}$	C	I	F	S	0.28 ± 0.02	0.65 ± 0.01	2300 ± 200	0.02
SIFH	$Open_{High}$	S	I	F	H	0.33 ± 0.03	1.11 ± 0.02	3400 ± 300	0.02
SIFS	$Open_{High}$	S	I	F	S	0.35 ± 0.04	1.04 ± 0.02	3000 ± 300	0.03

						L-Phenylalanine[c] K_M (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (M^{-1} s^{-1})	
WT	—	–	–	–	–	N.D.[f]	N.D.[f]	400 ± 100	
HEX	—	L	Y	I	L	0.27 ± 0.03	9.0 ± 0.2	33,000 ± 4000	
IYIT	Closed	I	Y	I	T	0.34 ± 0.03	10.1 ± 0.2	30,000 ± 3000	
VFIT	Closed	–	F	I	T	0.90 ± 0.06	22.8 ± 0.4	25,000 ± 2000	
VFIY	Closed	–	F	I	Y	0.42 ± 0.04	17.3 ± 0.4	41,000 ± 4000	
VYIT	Closed	–	Y	I	T	1.11 ± 0.07	41.8 ± 0.6	38,000 ± 2000	
VYIY	Closed	–	Y	I	Y	0.58 ± 0.02	20.9 ± 0.2	36,000 ± 1000	

continued on next page

Table 4.1 (continued)

		Mutations[b]				L-Aspartate[c]			
Enzyme[a]	Library	V35	K37	T43	N64	K_M (mM)	k_{cat} (s^{-1})	k_{cat}/K_M (M^{-1} s^{-1})	Selectivity[d]
IFCA	$Open_{Low}$	I	F	C	A	2.3 ± 0.2	47 ± 1	20,000 ± 2000	
MFCA	$Open_{Low}$	M	F	C	A	1.08 ± 0.08	35.9 ± 0.6	33,000 ± 3000	
VFCA	$Open_{Low}$	–	F	C	A	4.2 ± 0.2	47 ± 1	11,200 ± 600	
VFCS	$Open_{Low}$	–	F	C	S	12 ± 1	87 ± 3	7200 ± 700	
AIFS	$Open_{High}$	A	I	F	S	2.6 ± 0.2	0.54 ± 0.01	210 ± 20	
CIFC	$Open_{High}$	C	I	F	C	5.9 ± 0.4	3.89 ± 0.08	660 ± 50	
CIFS	$Open_{High}$	C	I	F	S	5.4 ± 0.4	0.257 ± 0.006	48 ± 4	
SIFH	$Open_{High}$	S	I	F	H	3.7 ± 0.3	0.245 ± 0.008	66 ± 6	
SIFS	$Open_{High}$	S	I	F	S	3.9 ± 0.2	0.294 ± 0.006	75 ± 4	

[a]Mutants are named on the basis of the amino-acid identity at the four designed positions (V35, K37, T43, N64). WT and HEX refer to wild-type AAT and the previously published hexamutant [46], respectively. [b]Mutations are numbered based on Uniprot sequence P00509. V35, K37, T43, and N64 correspond to V39, K41, T47, and N69 in the previously published crystal structures of wild-type AAT (PDB ID: 1ARS and 1ART). A dash (–) indicates the wild-type residue at that position. [c]All experiments were performed in triplicate using a single enzyme batch. Errors of regression fitting, which represent the absolute measure of the typical distance that each data point falls from the regression line, are provided. [d]Selectivity is defined as $(k_{cat}/K_{M,\text{L-phenylalanine}})/(k_{cat}/K_{M,\text{L-aspartate}})$. [e]HEX also contains the T104S and N285S mutations located near the cofactor phosphate moiety. [f]Individual K_M and k_{cat} values could not be determined accurately because saturation was not possible at the maximum substrate concentration tested (40 mM, the substrate's solubility limit). Catalytic efficiency was calculated from the slope of the linear portion ($[S] \ll K_M$) of the Michaelis–Menten model.

Table 4.2: Conformational equilibrium of AAT variants.

Enzyme	Computational[a]			Experimental[b]				
	E_{closed} (kcal mol^{-1})	E_{open} (kcal mol^{-1})	ΔE (kcal mol^{-1})	ΔH (kcal mol^{-1})	ΔS (kcal mol^{-1} K^{-1})	ΔC_p (kcal mol^{-1} K^{-1})	ΔG, 278 K (kcal mol^{-1})	ΔG, 303 K (kcal mol^{-1})
WT	−358.0	−372.2	14.2	N.D.	N.D.	N.D.	N.D.	N.D.
HEX	−350.2	−273.4	−76.8	−9.9	−0.032	0.972	−1.72	−0.31
VFIT	−366.1	−320.6	−45.5	−10.7	−0.038	0.667	−0.59	0.78
VFIY	−357.8	−300.0	−57.8	−19.6	−0.070	0.369	−0.38	1.59
VFCS	−385.6	−389.7	4.2	N.D.	N.D.	N.D.	N.D.	N.D.
AIFS	−237.2	−364.4	127.2	N.D.	N.D.	N.D.	N.D.	N.D.

[a]Boltzmann-weighted average potential energies ($T = 300$ K) for the open and closed state ensembles were computed using the Phoenix energy function for protein design (see Methods). Phoenix energies correspond to the sum of pairwise nonbonded interaction energies between rotamers and between rotamers and template in the folded state, without consideration of bonded interactions, entropy, or the energy of the unfolded state. The energy difference reported corresponds to $E_{\text{closed}} - E_{\text{open}}$. [b]Differences in enthalpy (ΔH), entropy (ΔS), heat capacity (ΔC_p), and Gibbs free energy (ΔG) are given in the direction of enzyme closing (e.g., $G_{\text{closed}} - G_{\text{open}}$), and were calculated using a reference temperature of 298 K. N.D., not determined.

Kinetic analysis of designs

We screened the three mutant libraries for transamination activity with the non-native substrate L-phenylalanine (see Methods) and selected the most active mutants from each library for kinetic analysis. All selected mutants catalyzed transamination of L-phenylalanine or L-aspartate with α-ketoglutarate, and displayed substrate inhibition with this acceptor substrate, as is the case for WT (Table 4.1). All Closed library mutants displayed catalytic efficiencies towards L-phenylalanine that were improved by approximately two orders of magnitude relative to WT (Table 4.1), comparable to HEX, and all were similarly or more active with this non-native substrate than with L-aspartate, in stark contrast with WT AAT, which prefers the native substrate by a factor of 100. Furthermore, all Closed library mutants were less catalytically efficient with L-aspartate than the WT despite having lower K_{M} values for this substrate, similar to HEX. These results are consistent with previous studies that showed that release of the oxaloacetate product resulting from L-aspartate transamination is dependent on a conformational change from the closed to the open state that is partially rate determining [92]. Meanwhile, $\mathrm{Open}_{\mathrm{Low}}$ and $\mathrm{Open}_{\mathrm{High}}$ mutants favored the native over the non-native substrate by up to 50-fold, similar to WT. $\mathrm{Open}_{\mathrm{Low}}$ mutants had lower K_{M} values and were more catalytically efficient than WT with both L-phenylalanine and L-aspartate, which could be due to the fact that these mutants have ΔE values similar to the WT but stabilize both the open and closed conformations by $>$10 kcal mol^{-1} (Table 4.2). This is not the case for $\mathrm{Open}_{\mathrm{High}}$ mutants, which have similar K_{M} values but are less catalytically efficient with the native substrate than WT, in agreement with our hypothesis that strong destabilization of the catalytically active closed conformation would result in less efficient catalysis. Overall, these kinetic results support the hypothesis that the change in substrate selectivity is linked to the conformational equilibrium previously suggested by the data from the HEX mutant [46, 47].

Structural analysis of designs

To provide structural information on the conformations adopted by AAT mutants, we turned to room-temperature X-ray crystallography, which provides insight into enzyme conformational ensembles under conditions that are relevant to catalysis [93] and free of potential distortions or conformational bias introduced by sample cryocooling [94]. We crystallized WT, HEX, and select variants from the Closed (VFIT and VFIY), Open_{Low} (VFCS), and $\text{Open}_{\text{High}}$ (AIFS) libraries. All six enzymes yielded crystals under similar conditions, which could only be obtained in the presence of maleate, an inhibitor that stabilizes the closed conformation when bound by both WT [95] and HEX [47]. To obtain structures in the absence of maleate, we applied a rigorous crystal soaking method to serially dilute and extract the inhibitor from the crystallized enzymes (see Methods). We collected X-ray diffraction data at room temperature (278 K) for all variants with the exception of AIFS, which could only be measured at cryogenic temperature (100 K) because we could only obtain small crystals that were not robust to radiation damage at non-cryogenic temperatures. We applied statistical criteria to assign high-resolution cut-offs of 1.37–2.31 Å for our data sets, and all structures were subsequently determined by molecular replacement in space group $P\,6_3$ with an enzyme homodimer in the asymmetric unit. Given the known impact of crystal packing and crystallization conditions on domain movement amplitude in AAT [95, 96], we confirmed that crystals of all variants with or without soaking had similar unit cell dimensions and identical space group. In the inhibitor-bound state, all structures were closed as expected due to electrostatic interactions between maleate and the side chains of Arg280 and Arg374. Upon soaking crystals of the WT enzyme to remove the bound inhibitor, we observed that one subunit (chain A) within the enzyme homodimer was in the open conformation (Fig. 4.3), confirming that the soaking procedure was able to remove the bound maleate, and that the crystal lattice could accommodate the domain rotation required for opening and closing of this subunit. Chain B however remained closed upon maleate removal, likely due to crystal packing interactions hindering domain rotation of this subunit and indicating that domain opening is not required for maleate removal. These

results are consistent with a previous study that showed that only one subunit of a homologous AAT closed upon soaking of the crystal in a substrate solution [96]. In the maleate-free structures, a sulfate ion and one or more ordered water molecules occupy the inhibitor binding site of both subunits, as was previously observed in WT [95] and HEX [47] structures at cryogenic temperatures.

Superposition of bound and unbound structures for each variant confirmed that closed library mutants VFIT and VFIY remained in the closed conformation in the inhibitor-free form, similar to HEX, while Open_{Low} variant VFCS and $\text{Open}_{\text{High}}$ variant AIFS adopted the open conformation, similar to WT (Fig. 4.3). AIFS is unique in that the electron density is especially weak in the regions corresponding to the helix formed by Pro12–Leu19 and the loop connecting moving and fixed domains (Leu31–Thr43), even at cryogenic temperatures, demonstrating that these structural segments are disordered when the enzyme adopts the open conformation. This result suggests that the four mutations of AIFS, three of which are found within the Leu31–Thr43 loop, contribute to destabilize the open conformation, consistent with the calculated E_{open} value of this variant (Table 4.2) being >7 kcal mol^{-1} higher than that of the other variants that favor the open state (WT and VFCS). Interestingly, the amplitude of the open/closed conformational transition that occurs in open variants upon maleate binding (Fig. 4.3) correlated with their computed ΔE values (VFCS $<$ WT $<$ AIFS), suggesting that this metric can be used to fine-tune this enzyme conformational landscape. Furthermore, DynDom analyses of all variants confirmed that only WT, VFCS, and AIFS undergo hinge motion upon maleate binding, which rotates the moving domain relative to the fixed domain by 4.6, 2.6, and 5.9 degrees, respectively.

NMR analysis of conformational landscapes

Having demonstrated crystallographically that our designed variants adopted the target conformation in the absence of ligand, we turned to NMR spectroscopy to gain insights into their conformational equilibria in solution, and compared results against those obtained for WT and HEX. We first measured ^{1}H-^{15}N HSQC spectra for WT in the presence and absence of L-aspartate. As expected for a

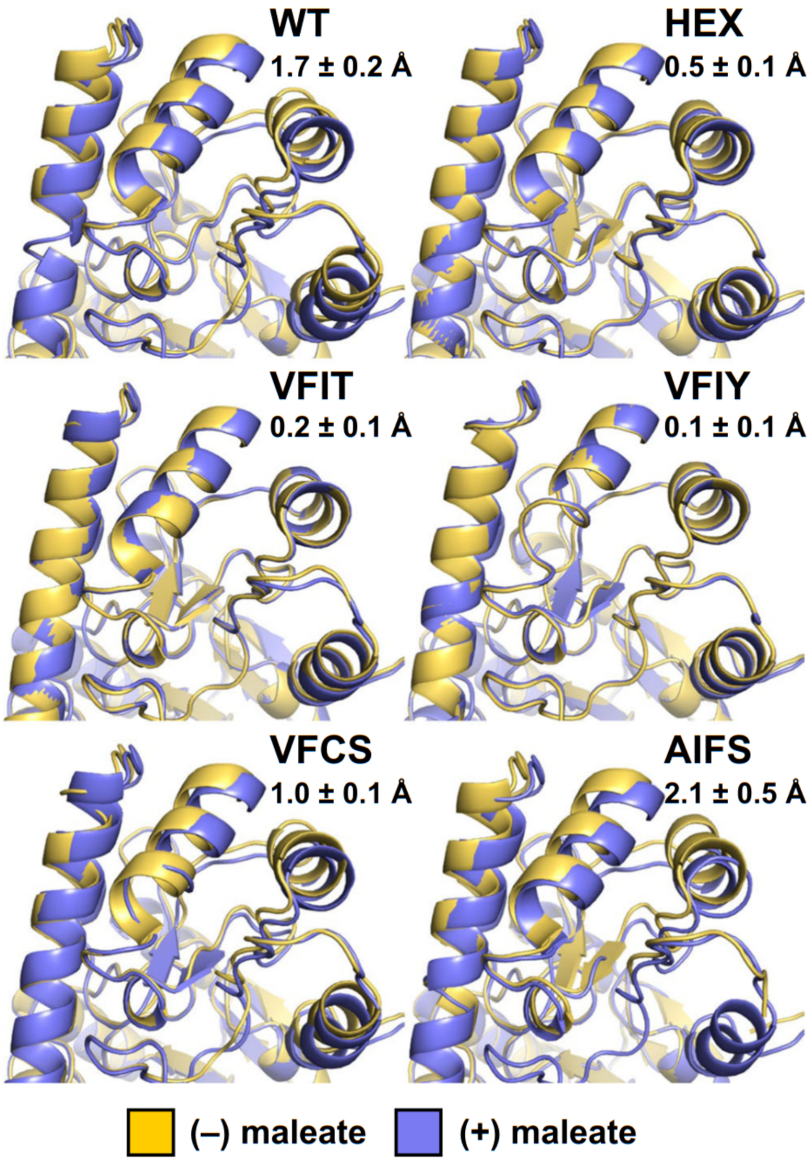


Figure 4.3: **Crystal structures.** Overlay of AAT structures (Chain A) in the presence and absence of maleate show transition from open to closed states upon inhibitor binding for WT, VFCS, and AIFS, but not for HEX, VFIT, and VFIY, which already adopt the closed conformation in the absence of bound inhibitor. Average displacements of helix formed by residues K355–F365 upon maleate binding are reported as the average pairwise distance of corresponding Cα atoms for the 11 residues comprising this helix (mean ± s.d.).

protein of this size, the spectrum consisted of a large number of peaks that were broad with a high degree of overlap. It was nonetheless possible to assign the unique chemical shifts of peaks from the indole NH group for three native Trp residues by comparison of these spectra with those acquired with single Trp mutants. This allowed assignment of a peak that was significantly broadened in spectra of both HEX and Closed library mutant VFIY to the indole NH from Trp307, whose side chain is closest to the designed hinge residues. Moreover, the Trp307 indole peak was no longer detectable when the L-aspartate substrate was present for both WT and mutant enzymes, most likely being broadened beyond detection. Since all conditions that favor the closed state (i.e., HEX and Closed library mutations and/or L-aspartate binding) broaden the Trp307 indole resonance, this exchange appears to be associated with the closed state, potentially due to conformational dynamics around the hinge.

In order to characterize the thermodynamics of the exchange processes around the hinge region, we labeled AAT variants at a single site with ^{19}F using site-specific incorporation of the noncanonical amino acid 4-trifluoromethyl-L-phenylalanine [97]. Phe217 was chosen as the incorporation site since it is proximal to hinge residues but is not in direct contact with the substrate (Fig. 4.4a). The ^{19}F spectrum of WT at 278 K showed a single peak centered at approximately -60 ppm that shifts downfield as temperature is increased (Fig. 4.4b), similar to what is observed for free 4-trifluoromethyl-L-phenylalanine in solution. By contrast, the HEX spectrum at 278 K showed a large broad peak centered at around -61.6 ppm, with another peak of substantially lower intensity also appearing at a similar shift to that observed in the WT spectrum (-60.4 ppm). The relative intensity of these 2 peaks changed as the temperature was increased, with the low-intensity peak increasing as the major peak decreased. This is characteristic of two-state exchange, with the equilibrium between the two states being shifted by the temperature change. Given that the crystal structure of HEX in its ligand-free form showed a closed conformation (Fig. 4.3), it is likely that the major peak reflects a local chemical environment created by the closed state, with a small population in an open state similar to that seen in the WT spectrum.

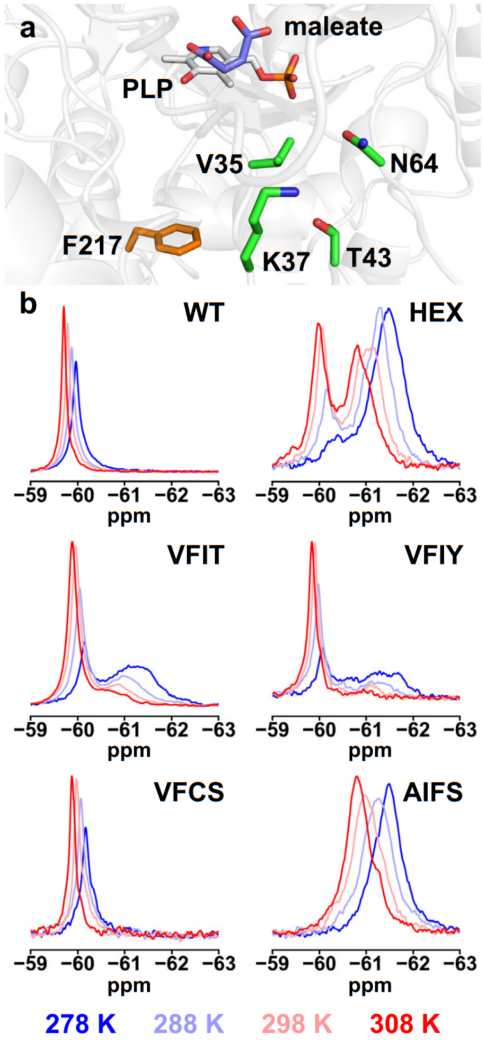


Figure 4.4: **Conformational landscape analysis by NMR. (a)** To evaluate conformational equilibrium of AAT variants, we introduced the fluorinated amino acid 4-trifluoromethyl-L-phenylalanine at position F217 (orange), which is located closer to hinge-bending and designed residues (green) than to the bound maleate inhibitor (blue). Crystal structure shown is that of wild-type (WT) AAT at 278 K (PDB ID: 8E9K). **(b)** ^{19}F NMR spectra of AAT variants in the absence of ligand show dynamic equilibrium between 278 K and 308 K for HEX, VFIT, VFIY, and AIFS, confirming that these proteins are undergoing exchange. This is not the case for WT and Open_{Low} library mutant VFCS, who both adopt predominantly the open conformation within this temperature range. For HEX and Closed library mutants VFIT and VFIY, the open conformation is enriched as temperature increases. For AIFS, spectra suggest that this $\text{Open}_{\text{High}}$ library mutant samples conformations distinct from those sampled by the other variants because it resembles neither the single peak open state nor the two-peak, open-closed states.

Using peak deconvolution and integration, it was possible to calculate relative populations for each species for variants undergoing the observed two-state exchange, along with the free energy difference between states. We calculated ΔG at 278 K for HEX to be -1.72 kcal mol^{-1} (Table 4.2), a small difference that would be compatible with the interconversion between these two states to be part of the catalytic cycle [98]. These peak volumes could also be calculated over the entire temperature range tested, giving rise to a van't Hoff relationship with a small degree of curvature. Deviations from linearity can occur when there are differences in heat capacity between the two states, as would be expected for a process involving a change in conformational states over the temperature range tested [99]. By contrast, the presence of a single peak in WT spectra suggests that the closed state is not significantly populated under these conditions, as expected from its open-state crystal structures obtained under both cryogenic [84, 95] and room-temperature conditions.

We next analyzed two mutants from the Closed library (VFIT and VFIY). ^{19}F NMR spectra at 278 K for both mutants showed a narrow peak centered at approximately -60.2 ppm that is similar to that of WT (Fig. 4.4). Spectra of these variants also showed another broad peak centered at approximately -61.2 ppm that is similar to the HEX peak characteristic of the closed conformation. The relative intensity of the two peaks showed similar temperature dependence, with an increase in the relative intensity of the WT-like peak as temperature was increased to 308 K. Peak deconvolution and integration were used to evaluate the population of the two states, and confirmed that at low temperatures both VFIT and VFIY favor the state resembling that adopted by the closed HEX mutant ($T < 288$ K or 283 K, respectively). However, unlike HEX, at higher temperatures the alternate conformation with the WT-like peak becomes the favored state (Table 4.2). Interestingly, ΔE values for these mutants were smaller in magnitude than that calculated for HEX, supporting the predictive nature of the calculated energy differences between open and closed states for these sequences.

To determine if these differences in the temperature dependence of exchange could be observed in the crystal state, we also solved the maleate-free structures

of WT, HEX, and VFIT at 303 K, and calculated isomorphous difference density maps by subtracting electron density at 278 K. Comparing VFIT data obtained at 278 K and 303 K results in substantial difference density throughout chain A, the chain that opens when WT crystals are soaked to remove maleate. By contrast, similar comparisons for WT and HEX showed relatively little difference density. This analysis confirms that VFIT undergoes larger local conformational changes than either HEX or WT when the temperature of the crystal is increased to 303 K, in agreement with our van't Hoff analysis of NMR data (Table 4.2). The agreement between temperature-dependent X-ray crystallography and NMR data provides strong evidence that the conformational exchange detected in the NMR experiments reflects a dynamic equilibrium between closed and open conformations, with mutations that favor the closed conformation also increasing selectivity toward L-phenylalanine.

We next analyzed two mutants designed to favor the open conformation in their ligand-free forms. Open_{Low} mutant VFCS showed ^{19}F NMR spectra that were very similar to those of WT within the tested temperature range (Fig. 4.4b), consistent with its preference for the open conformation (Fig. 4.3 and Table 4.2). However, $\text{Open}_{\text{High}}$ mutant AIFS gave rise to spectra that were distinct from those of all other variants (Fig. 4.4b), but could be deconvoluted to two exchanging peaks, to allow estimation of populations. We postulate that those peaks correspond to alternate open conformations distinct from the one sampled by the other variants. This hypothesis is supported by our observations that ligand-free AIFS is open at low temperature (Fig. 4.3) but contains disordered segments around the Pro12–Leu19 helix and the loop containing three of the four designed positions, which are located close to the Phe217 position where the ^{19}F label was introduced. The observed heterogeneity could therefore correspond to a mixture of these alternate open conformations. Additional support that the exchange detected in AIFS differed from that of HEX and closed mutants was provided by the van't Hoff analysis, which showed no curvature for AIFS unlike for the closed mutants, suggesting that the $\text{Open}_{\text{High}}$ variant does not undergo the open/closed conformational transition within this temperature range.

4.1.3 Discussion

Here, we successfully remodeled the conformational landscape of an enzyme via targeted alterations to the equilibrium between two distinct conformational states related by a hinge-bending motion. The resulting equilibrium shift promoted activity toward a non-native substrate, leading to a selectivity switch of up to 1900-fold. Our approach should be applicable to the redesign of other enzymes where alterations to known conformational equilibria, including those mediated by hinge-bending motions, are expected or hypothesized to increase catalytic efficiency [80, 100] or alter substrate selectivity [101]. As many enzymes undergo hinge-mediated domain motions during their catalytic cycles [102], the multistate design approach presented here should be straightforward to implement for such enzymes. Given the ability of our multistate design procedure to distinguish between closed and open states whose free energy difference is on the order of a single hydrogen bond, our approach could, in principle, also be applied to preferentially stabilize catalytically competent substates involving more subtle structural changes, such as backbone carbonyl flips [103] or side-chain rotations [104]. In such cases, it will be important to have structural information available for the target states. This methodology could therefore help to tailor catalytic efficiency or substrate selectivity by mimicking, *in silico*, the processes of evolution that harness altered conformational equilibria to tune function [80, 105, 106].

The predictive capacity of our multistate design framework could only be achieved by evaluating the energy of sequences on multiple conformational states. For example, the VFCS variant that prefers the open state is predicted to be more stable on the closed state than Closed library mutants VFIT and VFIY (Table 4.2), and more stable on the Open state than AIFS even though it is less open than this variant (Fig. 4.3). Furthermore, there were no obvious trends in designed mutations that could explain their effect on the conformational landscape, as none of these introduced bulkier or smaller amino acids at all or specific residue positions to cause or alleviate steric clashes in the open or closed conformations so as to shift the equilibrium towards one of these states, which has been the approach others have used to shift conformational equilibria [107]. Thus, subtle effects of

mutation combinations on the relative stability of each conformational state were likely responsible for the observed preference of mutants for the open or closed conformations.

Our results demonstrate the utility of multistate design, with ΔE values calculated from ensemble energies of open and closed states, for the targeted alteration of subtle conformational equilibria, an approach that represents a useful alternative to heuristic methods that others have used to tune the relative stability of protein conformational states [108]. Extending this concept, we envision that *de novo* design of artificial enzymes with native-like catalytic efficiency and selectivity for complex multistep chemical transformations will require a holistic approach where every conformational state and/or substate required to stabilize reaction intermediates and transition states are explicitly modeled, and their relative energies optimized. The multistate design method for conformational landscape remodeling presented here opens the door to alter this common type of enzyme conformational equilibrium to facilitate the creation of designer biocatalysts with tailored functionality.

4.1.4 Methods

Structure preparation and ensemble generation. Crystal structures of wild-type *Escherichia coli* AAT in its internal aldimine form (PDB ID: 1ARS [84]) or complexed with N-phosphopyridoxyl-L-glutamic acid (PDB ID: 1X28 [85]) were used to model the open and closed states, respectively. To eliminate biases that could arise during ensemble generation from the presence of substrate bound in the active site, we deleted coordinates for the N-phosphopyridoxyl-L-glutamic acid and catalytic K246 residue in the 1X28 structure and replaced them with coordinates for K246 and the pyridoxal 5′-phosphate (PLP) cofactor extracted from the 1ARS structure. These structures were then prepared for ensemble generation using the Molecular Operating Environment (MOE) software [109]. Hydrogens were added with the Protonate3D utility and manually adjusted to ensure that the protonation states of PLP and K246 were consistent with the aminotransferase catalytic mechanism [110]. The resulting structures were then solvated in a rectangular

box of water with counter ions (Na^+ and Cl^-) under periodic boundary conditions with a box cut-off of 6 Å, and energy minimized by conjugate gradient energy minimization to a root mean square gradient <7 kcal mol^{-1} $Å^{-1}$ using the AMBER99 force field [111] with a combined explicit solvent and implicit reaction field solvent model set up using the MOE software package. These structures were used as input templates to generate backbone ensembles with the PertMin algorithm [88, 89]. Briefly, two 50-member PertMin ensembles were created by randomly perturbing the coordinates of all heavy atoms of the two prepared AAT structures by ±0.001 Å along each Cartesian coordinate axis, and energy minimizing them using a truncated Newton [112] minimization algorithm for 100 iterations. The "Open" and "Closed" ensembles thus obtained displayed diversities (i.e. average backbone root mean square deviations between pairs of ensemble members) of 0.29 ± 0.03 Å and 0.32 ± 0.03 Å, respectively, and backbone root mean square deviations from the starting structure of 0.51 ± 0.02 Å.

Computational protein design. All calculations were performed using the Phoenix protein design software [91, 113] with the fast and accurate side-chain topology and energy refinement (FASTER) algorithm [114] for sequence optimization. The 2002 backbone-dependent Dunbrack rotamer library [115] with expansions of ±1 standard deviation around χ_1 and χ_2 was used to provide side-chain conformations of AAT residues to be threaded onto each fixed backbone template. Sequences were scored using the Phoenix energy function, a five-term potential energy function consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field [116] with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well depth of 8.0 kcal mol^{-1} and an equilibrium donor-acceptor distance of 2.8 Å [117], an electrostatic energy term modeled using Coulomb's law with a distance-dependent dielectric of 10, an occlusion-based solvation penalty term [91], and a secondary structural propensity term [118]. Sidechain rotamers of residues 35, 37, 43, and 64 were optimized on each backbone template using all proteinogenic amino acids with the exception of proline. Sidechain rotamers of residues within 5 Å of the designed residues were also optimized but their identities were not changed. The searched sequence space thus consisted

of 130,321 (19^4) sequences, resulting in >13 million individual sequence energies (130,321 sequences × 50 backbones × 2 ensembles). Boltzmann weighted average potential energies at 300 K were computed for each sequence on each ensemble, yielding energy values for the Closed and Open states (E_{closed} and E_{open}). Energy differences between conformational states ($\Delta E = E_{\text{closed}} - E_{\text{open}}$) were then computed for use in library design. To avoid favoring sequences displaying unfavorable E_{closed} and/or E_{open} values resulting in favorable ΔE (e.g., high, low, or high absolute ΔE values for Closed, Open_{Low}, or $\text{Open}_{\text{High}}$ libraries, respectively), which would be expected to be unstable, sequences whose E_{closed} and/or E_{open} value fell outside of the 75th percentile were discarded.

Library design. Library design was performed with the CLEARSS algorithm [90] using as input the ΔE values obtained as described above. For a specific library size configuration, which is the specific number of amino acids at each position in the protein (e.g., 4 amino acids at position 1, 3 amino acids at position 2, etc.), the highest probability set of amino acids at each position were included in the library. To identify the optimal library size configuration, all configurations that lead to a combinatorial library of target sizes (in this case, 20 ± 4 sequences) were scored by taking the sum of all partition functions of the chosen amino acid sets over all positions, and the highest scoring library was selected.

Mutagenesis. The wild-type *E. coli* AAT gene (Uniprot ID: P00509) with an N-terminal His-tag cloned into plasmid pET-45b (Novagen) via the NcoI/PacI restriction sites [119] was a generous gift from Michael D. Toney (University of California, Davis). Mutations were introduced into the AAT gene by overlap extension mutagenesis [120] using VentR DNA Polymerase. Briefly, external primers containing NdeI or BamHI restriction sites were used in combination with sets of complementary pairs of oligonucleotides containing mutated codons (individual codons for HEX and single point mutants, codon mixtures for the Closed, Open_{Low}, and $\text{Open}_{\text{High}}$ libraries) in individual polymerase chain reactions (PCRs). The resulting overlapping fragments were gel-purified (Omega Biotek) and recombined by overlap extension PCR. The resulting amplicons were digested with NdeI/BamHI, gel-purified, and ligated into the pET-11a expression vector (Novagen) with T4

DNA ligase. pBAD vectors (Invitrogen) harboring selected AAT genes flanked by NcoI/EcoRI restriction sites were prepared using a similar procedure. All constructs were verified by sequencing the entire open reading frame.

Preparation of clarified cell lysates. DNA libraries prepared as described above were transformed into chemically competent *E. coli* BL21-Gold (DE3) cells (Agilent). Colonies (180 per library) were picked into individual wells of V96 MicroWell polypropylene plates (Nunc) containing 300 μL of lysogeny broth (LB) supplemented with 100 μg mL^{-1} ampicillin and 10% glycerol. The plates were covered with a sterile breathable rayon membrane (VWR) and incubated overnight at 37 °C with shaking. After incubation, these mother plates were used to inoculate sterile Nunc V96 MicroWell polypropylene plates ("daughter" plates) containing 300 μL per well of Overnight Express Instant TB medium (Novagen) supplemented with ampicillin. Daughter plates were sealed with breathable membranes and incubated overnight (37 °C, 250 rpm). After incubation, cells were harvested by centrifugation (3000 × *g*, 30 min, 4 °C) and pellets were washed twice with phosphate-buffered saline (pH 7.4). Washed cell pellets were resuspended in lysis buffer (100 mM potassium phosphate buffer pH 8.0 containing 1× Bug Buster Protein Extraction Reagent [Novagen], 5 U mL^{-1} Benzonase Nuclease [EMD], and 1 mg mL^{-1} lysozyme). Clarified lysates were collected following centrifugation and stored at 4 °C until used in the screening assay.

Library screening. All assays were performed in 200-μL reactions at 37 °C in 100 mM potassium phosphate buffer (pH 8.0). The standard reaction mixture contained final concentrations of either 3 or 40 mM L-phenylalanine, 16 μM PLP, 0.2 mM α-ketoglutarate, 1 U of glutamate dehydrogenase (GDH) from bovine liver (Sigma), and 5 mM NAD^+. Plates containing the standard reaction mixture were incubated at 37 °C for 5 min prior to initiation of the reaction by addition of 10 μL of clarified cell lysates prepared as described above. Enzyme reactions were monitored by measuring absorbance of NADH at 340 nm every 12 sec for 30 or 60 min in individual wells of 96-well plates (Greiner Bio-One) using a SpectraMax 384 Plus plate reader (Molecular Devices). The four or five most active variants from each library were selected for further characterization.

Protein expression and purification. Selected mutants were expressed and purified as described by Mironov et al. [121]. Briefly, *E. coli* BL21-Gold (DE3) cells harboring expression vectors containing aminotransferase genes were grown at 37 °C in 500 mL LB medium supplemented with 100 μg mL^{-1} ampicillin until they reached an OD_{600} of 0.6. Isopropyl β-D-1-thiogalactopyranoside (1 mM) was added to the flasks to induce protein expression, followed by shaking overnight at 16 °C. Cells were harvested by centrifugation, resuspended in 10 mL lysis buffer (5 mM imidazole in 100 mM potassium phosphate buffer, pH 8.0), and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography using Ni–NTA agarose pre-equilibrated with lysis buffer in individual Econo-Pac gravity-flow columns (Bio-Rad). Columns were washed twice, first with 10 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0), and then with the same buffer containing 20 mM imidazole. Bound proteins were eluted with 250 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0) and exchanged into 100 mM sodium phosphate buffer (pH 8.0) using Econo-Pac 10DG desalting pre-packed gravity flow columns (Bio-Rad). For crystallography, proteins were further purified by gel filtration in 20 mM potassium phosphate buffer (pH 7.5) using an ENrich SEC 650 size-exclusion chromatography column (Bio-Rad). Purified samples were concentrated using Microsep Advance 10K centrifugal devices (Pall) to a final concentration of 170 μM. Protein concentrations were quantified using a modified version of the Bradford assay, where the calibration curve is constructed as a plot of the ratio of the absorbance measurements at 590 and 450 nm versus concentration [122].

Steady-state kinetics. To measure steady-state kinetics for the α-ketoglutarate acceptor substrate, assays were performed by varying the α-ketoglutarate concentration from 0.002 to 20 mM in the presence of 10–20 mM L-phenylalanine or L-aspartate, 5 mM NAD^+, 16 μM PLP, 1 U GDH, and approximately 10 mU of aminotransferase in 100 mM potassium phosphate buffer (pH 8, 37 °C). To measure steady-state kinetics for the L-aspartate and L-phenylalanine donor substrates, assays were performed by varying the amino-acid concentration from 0.002 to 40 mM in the presence of 0.1563–1.25 mM α-ketoglutarate, 5 mM NAD^+, 16 μM PLP,

1 U GDH, and approximately 10 mU of aminotransferase in 100 mM potassium phosphate buffer (pH 8, 37 °C). The pH of all reaction mixtures was adjusted to 8.0 prior to initiation of the reaction. Enzyme reactions were monitored by measuring absorbance of NADH at 340 nm (ε = 6220 M^{-1} cm^{-1}) every 12 sec for 30 or 60 min in individual wells of 96-well plates (Greiner Bio-One) using a SpectraMax 384 Plus plate reader (Molecular Devices). Path lengths for each well were calculated ratiometrically using the difference in absorbance of potassium phosphate buffer at 900 and 998 nm. Linear phases of kinetic traces were used to measure initial reaction rates. Initial reaction rates at different substrate concentrations were fit to the Michaelis–Menten equation using Python ver.2.7.15 with the `scipy.optimize.curve_fit` function (scipy ver.1.1.0). For mutant and substrate combinations resulting in substrate inhibition, fitting of the kinetic data was done with a rate equation that takes into account this type of inhibition: $v_0 = (v_{\max}[S])/(K_M + [S] + [S]^2/K_i)$.

Preparation of ^{15}N labeled proteins. Proteins for NMR spectroscopy were expressed using M9 minimal expression medium supplemented with 1 g L^{-1} ^{15}N-labeled ammonium chloride ($^{15}NH_4Cl$) for isotopic enrichment. Cultures were grown at 37 °C with shaking to an optical density at 600 nm of approximately 0.6, after which protein expression was initiated with 1 mM isopropyl β-D-1-thiogalactopyranoside. Following overnight incubation at 16 °C with shaking (275 rpm), cells were harvested by centrifugation and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography as described above, which was followed by gel filtration in 10 mM sodium phosphate buffer (pH 6.0) using an ENrich SEC 650 size exclusion chromatography column (Bio-Rad). Purified samples were concentrated using Amicon Ultracel-10K centrifugal filter units (EMD Millipore).

Preparation of ^{19}F site-specific labeled proteins. Site-specific incorporation of 4-trifluoromethyl-L-phenylalanine into selected AAT variants was prepared with a protocol adapted from Hammill et al. [97]. Briefly, chemically competent *E. coli* DH10B cells (Thermo Fisher) harboring the pDule-4-tfmF A65V S158A plasmid (Addgene plasmid #85484) [123], which encodes an orthogonal

aminoacyl-tRNA synthetase and cognate amber suppressing tRNA for site-specific incorporation of 4-trifluoromethyl-L-phenylalanine, were transformed with pBad vectors (Invitrogen) containing mutated genes of AAT variants in which an amber stop codon was introduced at residue position 217 to allow direct incorporation of the fluorinated amino acid. Labeled AAT variants were expressed in 2 L of LB containing 100 μg mL^{-1} ampicillin and 15 μg mL^{-1} tetracycline at 37 °C. After cells were grown for 1 h, 468 mg of 4-trifluoromethyl-L-phenylalanine (SynQuest Laboratories) was added to the flask to give a final concentration of 1 mM. Once the cell culture reached an OD_{600} of 0.6, L-arabinose was added to a final concentration of 0.2% to induce protein expression. Following overnight incubation at 16 °C with shaking, cells were harvested by centrifugation, resuspended in 10 mL lysis buffer, and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were then extracted and purified by immobilized metal affinity chromatography, as described above. Elution fractions containing the aminotransferases were concentrated through centrifugation (Pall Microsep Advance Centrifugal Device 10K), resuspended in 10 mM potassium phosphate buffer pH 6.0 and further purified through size exclusion chromatography, as described above. Elution fractions containing the purified aminotransferase were combined and concentrated through centrifugation.

^{1}H-^{15}N heteronuclear single quantum coherence spectroscopy. ^{15}N-labeled AAT samples for NMR (wild type, HEX, VFIY, and single point mutants Trp124Phe, Trp194Phe, Trp307Phe) consisted of 0.2–0.6 mM protein in 10 mM sodium phosphate buffer (pH 7.4), 10 μM EDTA, 0.02% sodium azide, and 10% D_2O. HSQC experiments were performed on a Bruker AVANCEIII HD 600 MHz spectrometer equipped with a triple resonance cryoprobe. 128 scans were accumulated to acquire each spectrum.

^{19}F nuclear magnetic resonance spectroscopy. Protein samples used for ^{19}F NMR analysis were diluted to concentrations of 75–280 μM in 500 μL of 10 mM potassium phosphate buffer (pH 6.0), 100 μM EDTA, 0.02% sodium azide, and 10% D_2O. All ^{19}F NMR spectra were acquired with a Bruker Avance 500 MHz spectrometer with 10 sec acquisition delays. 512 scans were accumulated for one-

dimensional ^{19}F chemical shift analysis at 5, 10, 15, 20, 25, 30, and 35 °C. Data were processed with an exponential window function (20 Hz linebroadening) using TopSpin 3.6.1 (Bruker Biospin). The ^{19}F NMR spectra of fluorinated HEX, VFIT, and VFIY showed two resonances, which we interpret to correspond to two distinct conformations that do not rapidly exchange with each other on the ^{19}F-NMR timescale. All spectra were deconvoluted into two Lorentzian curves using Python v3.7.9 with the LMFIT package for nonlinear least-squares minimization and curve-fitting [124]. The resulting two curves were integrated to measure the relative populations of each conformational state. The ratio of the two populations were used to calculate equilibrium constants at each temperature and fit using the Solver function of Excel to the nonlinear van't Hoff equation [125] (Equation 4.1), where enthalpy and entropy are not assumed to be temperature-independent (i.e. $\Delta C_P \neq 0$) to extract thermodynamic parameters of equilibrium. When fitting to the linear van't Hoff equation, ΔC_P was simply set to zero.

$$\ln K_{\mathrm{eq}} = -\frac{\Delta H^{\circ}_{\mathrm{ref}}}{R}\left(\frac{1}{T}\right) + \frac{\Delta S^{\circ}_{\mathrm{ref}}}{R} - \frac{\Delta C_P}{R}\left(\frac{T - T_{\mathrm{ref}}}{T} + \ln\frac{T_{\mathrm{ref}}}{T}\right) \tag{4.1}$$

where T_{ref} is an arbitrary reference temperature (set to 298 K), and $\Delta H^{\circ}_{\mathrm{ref}}$ and $\Delta S^{\circ}_{\mathrm{ref}}$ are $\Delta H^{\circ}(T)$ and $\Delta S^{\circ}(T)$ evaluated at T_{ref}, respectively.

Crystallization. Purified AAT variants were prepared in 20 mM potassium phosphate buffer (pH 7.5) with 2 mM EDTA and 10 μM PLP. Maleic acid was dissolved in water to make a 1 M stock solution, and then added to each protein solution yielding a final maleate concentration of 20 mM. For each enzyme variant, we carried out initial crystallization trials in 15-well hanging drop format using EasyXtal crystallization plates (NeXtal) and a crystallization screen that was designed to explore parameter space around the crystallization conditions reported by Islam et al. [85]. Crystallization drops were prepared by mixing 1 μL of protein solution with 1 μL of the mother liquor and sealing the drop inside a reservoir containing 500 μL of mother liquor. The mother liquor solutions contained ammonium sulfate as a precipitant. For all six enzymes, a microseeding protocol was required to obtain high-quality crystals. Microseeds were prepared by crushing initial crystals in their mother liquor using a glass rod, and were subsequently

streaked into the crystallization drops using a cat whisker. Because maleate was required for crystallization, ligand-free structures were obtained by soaking the crystal in new drops of 1 μL mother liquor containing 10 μM PLP but no maleate, allowing maleate in the crystal to diffuse out. Each crystal was treated in this way six subsequent times, 12 hours apart, to achieve removal of bound maleate.

X-ray data collection and processing. Prior to X-ray data collection, crystals were mounted on polymer MicroMounts (MiTeGen) and sealed using a MicroRT tubing kit (MiTeGen). Single-crystal X-ray diffraction data was collected on beamline 8.3.1 at the Advanced Light Source. The beamline was equipped with a Pilatus3 S 6M detector and was operated at a photon energy of 11,111 eV. Crystals were maintained at either 100 K, 278 K, or 303 K throughout the course of data collection. Each data set was collected using a total X-ray dose of 50–100 kGy and covered a 180° wedge of reciprocal space. Multiple data sets were collected for each enzyme variant either from different crystals, or if their size permitted, from unique regions of single large crystals. X-ray data were processed with the Xia2 program [65], which performed indexing and integration with DIALS [64], followed by scaling with DIALS.SCALE [126]. The resolution cut-off was taken where the $CC_{1/2}$ and $\langle I/\sigma I \rangle$ values for the intensities fell to approximately 0.5 and 1.0 respectively.

Structure determination. We obtained initial phase information for calculation of electron density maps by molecular replacement using the program Phaser [127], as implemented in v1.17.1.3660 of the PHENIX suite [128], with the crystal structure of wild-type *Escherichia coli* AAT complexed with N-phosphopyridoxyl-L-glutamic acid (PDB ID: 1X28 [85]) as search model. All AAT variants crystallized in the same crystal form, containing two chains of the molecule in the crystallographic asymmetric unit. Next, we rebuilt the initial model using the electron density maps calculated from molecular replacement. We then performed additional, iterative refinement of atomic positions, individual atomic displacement parameters (B-factors), and occupancies using a translation-libration-screw (TLS) model, a riding hydrogen model, and automatic weight optimization, until the model reached convergence. All model building was performed using Coot 0.8.9.2

[68] and refinement steps were performed with `phenix.refine` (v1.17.1.3660) within the PHENIX suite [128, 66]. Restraints for PLP in its internal aldimine form were generated using `phenix.elbow` [129], starting from coordinates available in the Protein Data Bank [130] (PDB ligand ID: PLP), and manually edited to ensure planarity of the pyridine ring and correct geometry of the imine bond between PLP and K246.

Per-chapter acknowledgment

The text and figures of Part A of this chapter are adapted from St-Jacques, A. D.; **Rodriguez, J. M.**; Eason, M. G.; Foster, S. M.; Khan, S. T.; Damry, A. M.; Goto, N. K.; Thompson, M. C.; Chica, R. A. "Computational remodeling of an enzyme conformational landscape for altered substrate selectivity." *Nature Communications* **14**, 6058 (2023). `https://doi.org/10.1038/s41467-023-41762-0`. This work is distributed under a Creative Commons Attribution 4.0 International License (`https://creativecommons.org/licenses/by/4.0/`). The dissertation author was a co-author of this publication and was responsible for crystallizing the AAT variants, performing the X-ray diffraction experiments, and contributing to structure refinement.

4.2 Part B: Design of efficient artificial enzymes using crystallographically enhanced conformational sampling

Abstract

The ability to create efficient artificial enzymes for any chemical reaction is of great interest. Here, we describe a computational design method for increasing the catalytic efficiency of *de novo* enzymes by several orders of magnitude without relying on directed evolution and high-throughput screening. Using structural ensembles generated from dynamics-based refinement against X-ray diffraction data

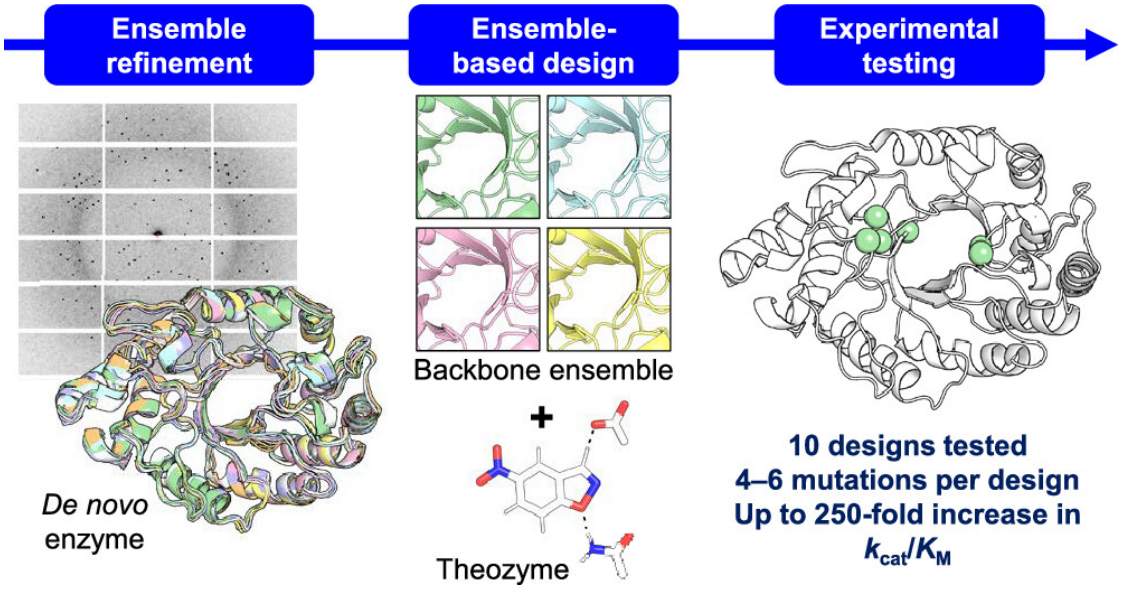


Figure 4.5: Graphical abstract from the original publication.

collected from crystals of Kemp eliminases HG3 (k_{cat}/K_M 125 M^{-1} s^{-1}) and KE70 (k_{cat}/K_M 57 M^{-1} s^{-1}), we design from each enzyme $\leq$10 sequences predicted to catalyze this reaction more efficiently. The most active designs display k_{cat}/K_M values improved by 100–250-fold, comparable to mutants obtained after screening thousands of variants in multiple rounds of directed evolution. Crystal structures show excellent agreement with computational models, with catalytic contacts present as designed and transition-state root-mean-square deviations of $\leq$0.65 Å. Our work shows how ensemble-based design can generate efficient artificial enzymes by exploiting the true conformational ensemble to design improved active sites.

4.2.1 Introduction

The ability to create efficient artificial enzymes for any desired chemical reaction is of great interest as it would open the door to a broad range of applications in industry and medicine. Algorithms for *de novo* enzyme design [131, 132] can be used to create artificial enzymes for reactions for which no naturally occurring enzyme is known. While this computational design process has been used to produce *de novo* enzymes for various model organic transformations [133, 134, 135, 136, 113], downstream directed evolution has been necessary to convert these enzyme prototypes

into catalysts as efficient as their natural counterparts [137, 138, 139, 140, 141, 142]. However, directed evolution is time-consuming and impractical for many enzymatic reactions of interest, for which high-throughput screening assays are unavailable. Therefore, a computational design procedure to enhance the catalytic efficiency of *de novo* enzymes would be beneficial as it could bypass directed evolution and accelerate the creation of efficient artificial enzymes.

Structural analyses of high-activity variants evolved from *de novo* enzymes revealed several mechanisms by which catalytic functions can be augmented. These mechanisms include optimization of hydrogen-bond geometries between catalytic residues and the transition state (TS) [137], enhancement of active-site complementarity to the TS [137, 138, 143], improvement of active-site preorganization [143], and enrichment of catalytically competent substates in the conformational ensemble [103, 78, 144]. These structural effects are challenging to design using current enzyme design protocols [133, 113, 138] because they are not sufficiently precise [137] and do not consider the dynamic nature of enzyme structures [103]. Recently, new methodologies have been developed that take advantage of the spatiotemporal averaging inherent to X-ray crystallography—crystallographic data represent billions of crystallized molecules that are all sampling different conformations during data collection—to model the conformational ensembles of proteins with high fidelity [145, 146]. New design strategies that could exploit this experimental information should facilitate the design of efficient artificial enzymes by enabling the accurate modeling of catalytically productive conformational substates.

Here, we report a computational procedure for enhancing the catalytic efficiency of *de novo* enzymes that exploits crystallographically derived ensemble models for increased accuracy. Starting from the noncryogenic crystal structure of *de novo* Kemp eliminase HG3 [113, 143], we generate a structural ensemble fitted to the diffraction data [145, 147] and use it to design mutant sequences predicted to efficiently catalyze the Kemp elimination. We filter these sequences using a variety of structural criteria aiming to reproduce molecular mechanisms known to augment enzyme catalysis and experimentally characterize 10 sequences. The most active design displays a catalytic efficiency improved by approximately 250-fold, a result

comparable to those obtained by seven rounds of directed evolution and screening of >7000 variants [137]. We solve crystal structures of several high-activity variants that show excellent agreement with the computational models and confirm that catalytic efficiency is enhanced via the intended molecular mechanisms. We also apply our approach to another *de novo* Kemp eliminase, KE70 [133], which yields comparable results. Our work shows how ensemble-based design can generate efficient artificial enzymes by exploiting information derived from X-ray crystallography to identify productive substates for efficient catalysis.

4.2.2 Results

Ensemble-based enzyme design

To create efficient *de novo* enzymes without relying on iterative rounds of mutagenesis and high-throughput screening, we implemented an ensemble-based computational enzyme design pipeline that uses protein backbones generated by ensemble refinement of diffraction data [145] as templates for design (Fig. 4.7). Because the diffraction data and calculated electron density maps are derived from crystals that contain billions of molecules, which are all sampling different conformational states throughout data collection, the resulting ensemble models represent a distribution of conformations that are well populated in the ensemble at equilibrium [148]. Thus, we postulated that performing crystallographic ensemble refinement on a low activity *de novo* enzyme would allow sampling of conformational substates that can more accurately accommodate the TS and its interactions with catalytic residues than the single conformation modeled using traditional refinement methods, thereby facilitating the design of improved active sites. To test our hypothesis, we applied an ensemble-based design to increase catalytic efficiency of the *de novo* enzyme HG3 [113] that catalyzes the Kemp elimination (Scheme 1), a well-established model organic transformation for benchmarking computational enzyme design methods. We selected HG3 as the case study for three reasons. First, X-ray diffraction data collected at noncryogenic temperatures (277 K) were available [143], which is ideal for ensemble refinement because cryocooling is known to narrow the conformational ensemble of crystallized proteins [94], reducing the

number of conformations that are represented in the ensemble-averaged data. Second, these data were collected from crystals with a bound TS analogue that is resolved in the electron density [143], allowing inclusion of the TS in the ensemble models. Finally, HG3 was subjected to an extensive directed evolution campaign that increased catalytic efficiency by almost 3 orders of magnitude [137], allowing us to compare results with those from purely experimental approaches.

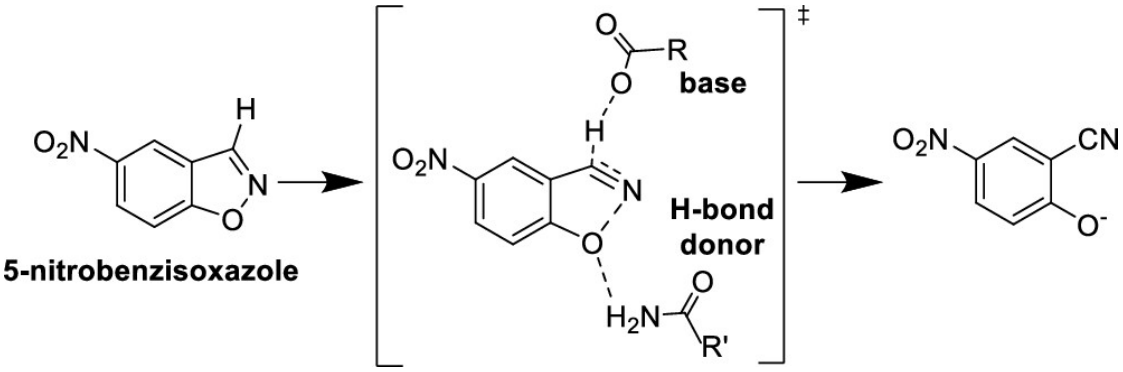


Figure 4.6: **Scheme 1.** Reaction catalyzed by HG-series Kemp eliminases.

We generated an ensemble of 84 backbone templates from the crystal structure of HG3 bound with the 6-nitrobenzotriazole (6NT) TS analogue (PDB ID: 5RGA [143]) and used these structures for placement of four theozymes containing an Asp residue as catalytic base, and either Gln, Asn, Ser, or Thr (Fig. 4.7a) to stabilize the binding of the substrate and properly orient it for reaction with the base. We kept the base position fixed to that of HG3 (D127) but allowed the algorithm to place the hydrogen-bond donor at multiple positions in the active site given that no amino acid residue in HG3 forms this catalytic contact with 6NT [143]. We then repacked the active site around the most stable theozyme placed on each backbone template by optimizing the identity and side-chain conformation of residues within 4 Å of the TS and catalytic residues to maximize beneficial nonbonding interactions with the theozyme, which has been shown to contribute to increased catalytic activity [137].

After repacking, we filtered designs using a series of structural metrics aiming to reproduce mechanisms known to augment enzyme catalysis (Fig. 4.7b). First, we measured angles defining the hydrogen bond between D127 and the TS for the top-

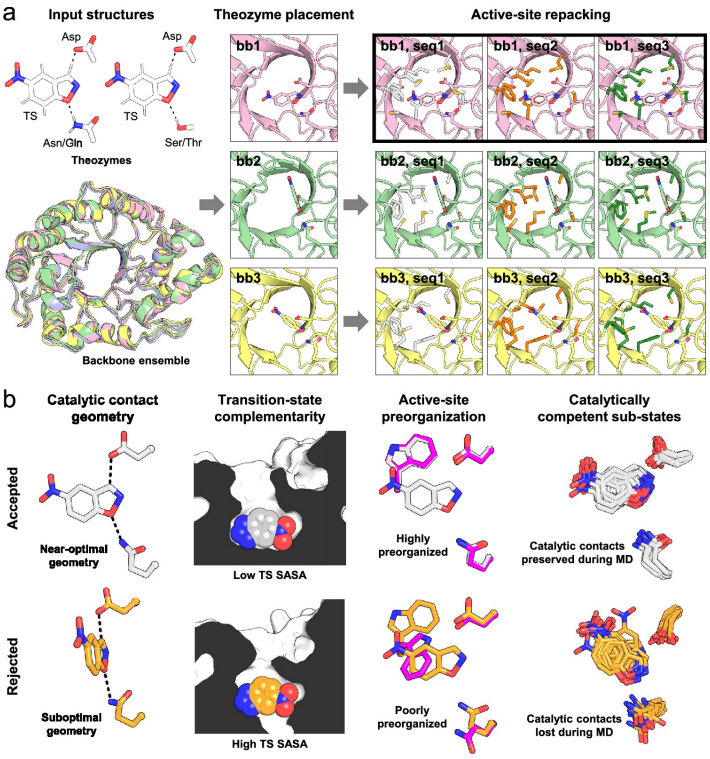


Figure 4.7: **Ensemble-based computational enzyme design. (a)** This procedure requires as inputs one or more theozymes, which are combinations of catalytic groups arranged in a unique geometry to stabilize a TS, and a backbone ensemble that approximates the intrinsic flexibility of the target protein fold. The algorithm proceeds in two stages: placement of each theozyme onto every member of the backbone ensemble (bb1–3), followed by repacking of active-site residues to identify sequences (seq1–3) that can further stabilize the TS. During repacking, both the identity and conformation of designed active-site residues are allowed to vary, and the TS is rotated and translated to reduce steric clashes with repacked side chains. In this toy example, only the theozyme containing a Gln hydrogen-bond donor is shown for the sake of clarity. After active-site repacking, designs from single theozyme/template combinations are selected based on energy (bold outline) and filtered using structural criteria. **(b)** Filtering of designs is performed to reproduce structural features that have previously been associated with efficient catalysis in other Kemp eliminases. Accepted designs are those that have (1) near-ideal contact geometries between the TS and catalytic side chains (dashed lines), (2) active sites with high structural complementarity to the TS, as calculated by low TS solvent-accessible surface area (SASA), (3) highly preorganized active sites with most designed residues predicted to adopt the same rotameric configuration in the presence (white/orange) and absence (magenta) of TS, and (4) enriched catalytically competent substates, which are snapshots that preserve catalytic contacts over the course of short molecular dynamics (MD) simulations.

scoring design obtained on each backbone template, and selected sequences from the single backbone template that gave the most stable design with angles that were closest to the optimal geometries calculated for acetamide dimers [149], which mimic the geometry needed for efficient proton transfer. The selected backbone, which we name bb05, yielded sequences that contain the Q50 hydrogen-bond donor that emerged during evolution of HG3 and that is present in the highest activity variants of this enzyme family [137]. Second, we evaluated active-site complementarity to the TS by calculating the solvent-accessible surface area of this ligand in the computational models generated for 1000 sequences designed on backbone bb05. Only designs where $\geq$95% of the TS surface is buried were kept. Third, we evaluated active-site preorganization by calculating the percentage of designed active-site residues that were predicted to adopt similar rotamers in the absence and presence of the TS. Sequences whose active sites were predicted to be >70% preorganized were kept. We also eliminated sequences with calculated energies over 10 kcal mol^{-1} higher than that of an HG3 single point mutant containing the D127/Q50 catalytic dyad as well as those that contained more than four methionines or more than one cysteine in the active site to avoid excessive conformational heterogeneity and prevent formation of disulfide bonds, respectively. Importantly, we did not eliminate entirely these amino acids from our designed active sites nor explicitly specify which positions should contain the methionines or cysteine that were preserved. This filtering process narrowed the original list of 1000 designs obtained from backbone bb05 down to 55 designs. A final filtering step consisting of a 20 ns MD simulation was used to identify designs that were enriched in catalytically competent conformational substates, which we define as snapshots that make the designed catalytic contacts. This final filtering resulted in nine sequences containing four or five mutations (Table 4.3), which we experimentally characterized.

Kinetic characterization of designs

All nine designs catalyzed the Kemp elimination, with all but two having catalytic efficiencies greater than that of HG3 (Table 4.3). The most active of these

variants, HG630, had a catalytic efficiency of 15,000 ± 500 M^{-1} s^{-1}, a value that is 120-fold higher than that of the starting template HG3. Furthermore, the three most active designs (HG198, HG259, and HG630) were more catalytically efficient than the HG3.3b variant obtained after three rounds of directed evolution and screening of >3000 mutants [137]. These designs all contain the M84C mutation that was previously shown to increase catalytic efficiency by 1 order of magnitude when combined with K50Q to yield the HG3-K50Q/M84C double mutant [103] (Table 4.3). Comparison of the sequences of HG44 and HG57, which differ by a single point mutation at position 90, suggested that the introduction of the Q90F mutation into HG630 would further enhance its catalytic efficiency as it doubles k_{cat}/K_M when introduced into HG57 to yield HG44. Introduction of this mutation into HG630 doubled its catalytic efficiency and resulted in a hexamutant (HG649) that is 256-fold more catalytically efficient than HG3 (k_{cat}/K_M = 32,000 ± 3000 M^{-1} s^{-1}) and more active than HG3.7, a variant found after seven rounds of directed evolution involving the screening of >7000 mutants [137] (Table 4.3, Fig. 4.8a). HG649 shares a majority of its mutation sites with HG3.7, but the combination of specific amino acid substitutions is distinct with only two mutations in common (Fig. 4.8b). These results demonstrate how ensemble-based enzyme design produces similar activity increases as directed evolution but without extensive experimental screening. Interestingly, HG649 was one of the 55 designs that passed all filtering criteria except for the final MD analysis, which led to its elimination due to its low predicted catalytic competency. This false negative, along with the two false positives (Table 4.3, HG378 and HG57), indicates that improvements are required to make the MD filtering more robust, for example, by running multiple replicas and/or longer simulations, and suggests that other eliminated sequences may also display high catalytic efficiency.

Table 4.3: Kinetic parameters of Kemp eliminases.

Enzyme	Mutations[a]	# mutations[a]	k_{cat} (s^{-1})[b]	k_{cat}/K_M (M^{-1} s^{-1})[b]	Relative k_{cat}/K_M
HG series — controls					
HG3	—	—	N.D.	125 ± 8	1
HG3.3b	V6I K50H M84C S89R Q90D A125N	6	N.D.	3400 ± 700	27
HG3-K50Q/M84C	K50Q M84C	2	N.D.	4300 ± 300	34
HG3.7	V6I Q37K K50Q M84C S89R Q90H A125N	7	N.D.	24,000 ± 3000	192
HG series — designs					
HG378	A21C K50Q Q90L A125V W275V	5	N.D.	70 ± 5	0.6
HG57	A21C K50Q Q90L A125V	4	N.D.	120 ± 8	1
HG44	A21C K50Q Q90F A125V	4	N.D.	250 ± 10	2
HG603	K50Q Q90F A125V W275V	4	N.D.	800 ± 10	6.4
HG493	K50Q Q90F A125T W275T	4	N.D.	2100 ± 100	17
HG185	A21C K50Q Q90F A125T W275V	5	N.D.	2300 ± 100	18
HG198	K50Q M84C A125T F267M	4	N.D.	8000 ± 300	64
HG259	K50Q M84C Q90L A125T F267M	5	N.D.	13,000 ± 400	104
HG630	K50Q M84C A125T F267M W275V	5	N.D.	15,000 ± 500	120
Without MD filtering[d]					
HG649	K50Q M84C Q90F A125T F267M W275V	6	N.D.	32,000 ± 3000	256
KE70 series — controls					
KE70	—	—	N.D.	57 ± 3	1
KE70-S20a/A240a[e]	S20a A240a	2	N.D.	150 ± 7	2.6
R2 7/12F	D23G Y48F D212E H251Y	4	0.50 ± 0.01	500 ± 30	9
R4 8/5A[e]	S20a K29N Y48F W72C H166Y K197N A204V G240a	8	3.0 ± 0.1	3000 ± 300	53
KE70 series — designs[f]					
KE701[e]	H16D S20a D45Q Y48F W72C G101S S138Q H166N V168A A204V A240a	11	0.040 ± 0.001	400 ± 80	7
KE703[e]	H16D S20a D45N Y48F W72V G101M S138Q H166N V168A A204V A240a	11	0.50 ± 0.01	2500 ± 300	44
KE703b[e]	S20a Y48F W72V G101M S138Q H166N V168A A204V A240a	9	N.D.	5500 ± 100	96

[a]Mutations are relative to the starting template (HG3 or KE70). [b]Data represent the average of six to 12 individual replicate measurements from two to four independent protein batches. N.D. indicates that individual parameters K_M and k_{cat} could not be determined accurately because saturation was not possible at the maximum substrate concentration tested (2 mM), which is the substrate's solubility limit. In these cases, catalytic efficiencies (k_{cat}/K_M) were calculated from the slope of the linear portion ($[S] \ll K_M$) of the Michaelis–Menten model ($v_0 = (k_{cat}/K_M)[E_0][S]$). Errors of regression fitting are provided. [d]HG649 is one of the 55 designs that passed all filtering criteria prior to the final MD step. [e]These enzymes contain two insertions, S20a and A240a/G240a. [f]The KE702 design was inactive.

Crystal structures

To evaluate whether the catalytic efficiency of our designs was increased via the predicted molecular mechanisms (Fig. 4.7b), we solved room-temperature (280 K) X-ray crystal structures of four of the most active designs (HG185, HG198, HG630, and HG649) in the presence and absence of the 6NT TS analogue. We determined structures at room temperature for direct comparison with the HG3 noncryogenic structure used to generate the ensemble and because this method can reveal conformational heterogeneity in protein structures that would not be visible at cryogenic temperatures, providing insights into the conformational ensemble that is sampled by a protein [148, 150]. Crystals of HG185, HG630, and HG649 were obtained in the absence of 6NT and soaked into solutions containing this molecule prior to X-ray data collection to obtain the bound structures. These three enzyme variants all crystallize in space group $P2_1$ with two molecules in the asymmetric unit, roughly matching the crystal form reported for HG3 [143]. On the other hand, HG198 could only be crystallized in the presence of 6NT, and its unbound structure was derived from serially soaking these crystals in solutions lacking the TS analogue prior to data collection. In the presence of 6NT, HG198 crystallizes in low symmetry space group $P1$ with three molecules in the asymmetric unit, but removing the TS analogue via serial soaking transforms the crystals to space group $P2_1$ with a single molecule in the asymmetric unit. This unusual transformation of the crystal system was consistent over multiple crystal specimens. All crystals diffracted at resolutions of 1.32–1.90 Å, which enabled us to create high-quality structural models.

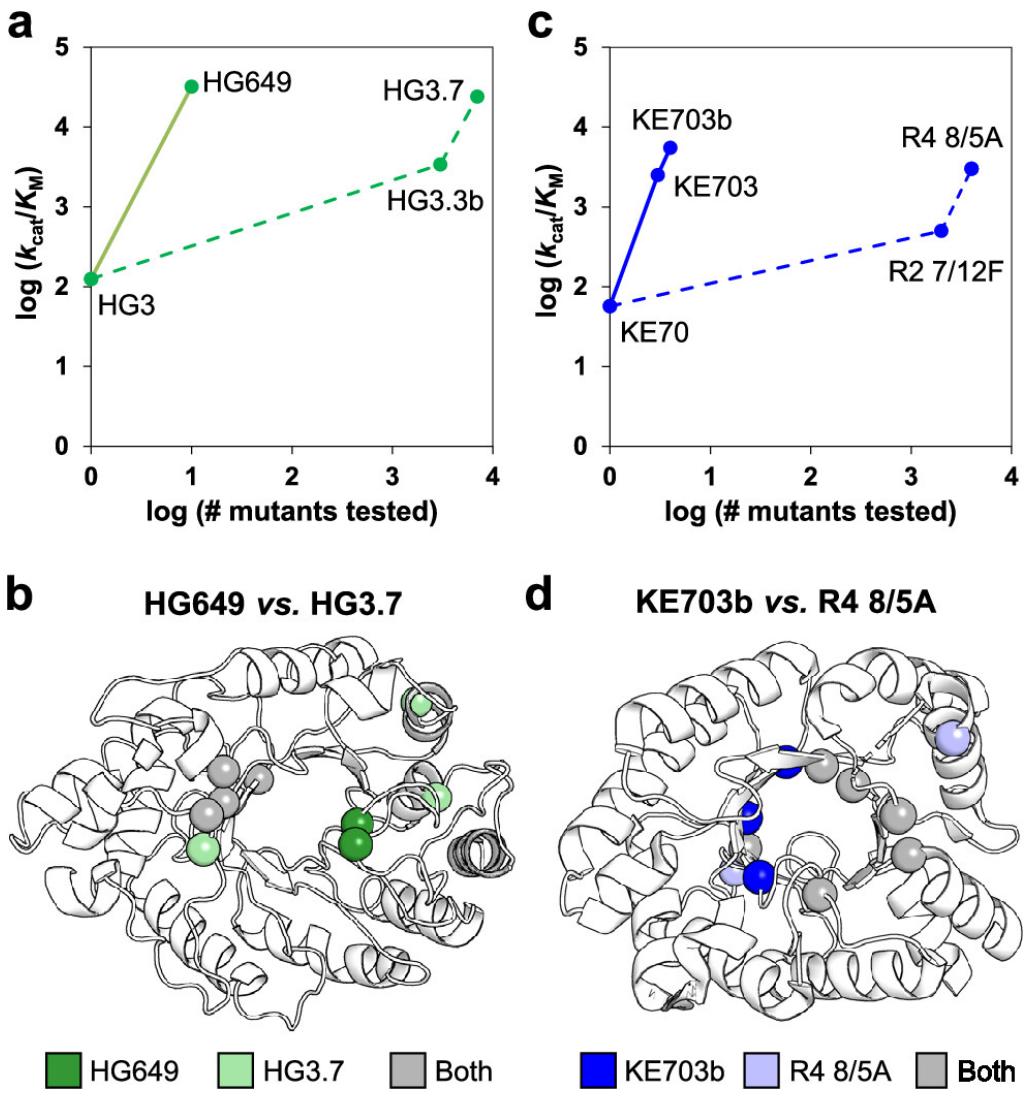


Figure 4.8: **Ensemble-based design vs directed evolution.** **(a)** Ensemble-based design yields catalytic efficiency improvements comparable to those achieved by directed evolution with a fraction of the experimental screening effort. Full and dashed lines show ensemble-based design and directed evolution trajectories, respectively. **(b)** Ensemble-based design identifies a majority of the mutation sites (spheres) found by directed evolution (4/6) but mutation combinations of designed (HG649) and evolved (HG3.7) enzymes are distinct as these share only two mutations (K50Q and M84C). **(c)** Full and dashed lines show ensemble-based design and directed evolution trajectories, respectively. KE703b differs from KE703 by reversion of its D16/N45 residues to the H16/D45 catalytic dyad found in KE70 and its evolved variants. **(d)** Designed variant KE703b shares 6 of the 9 mutation sites (spheres) found by directed evolution but has only three mutations in common with evolved variant R4 8/5A (S20a, Y48F, A204V).

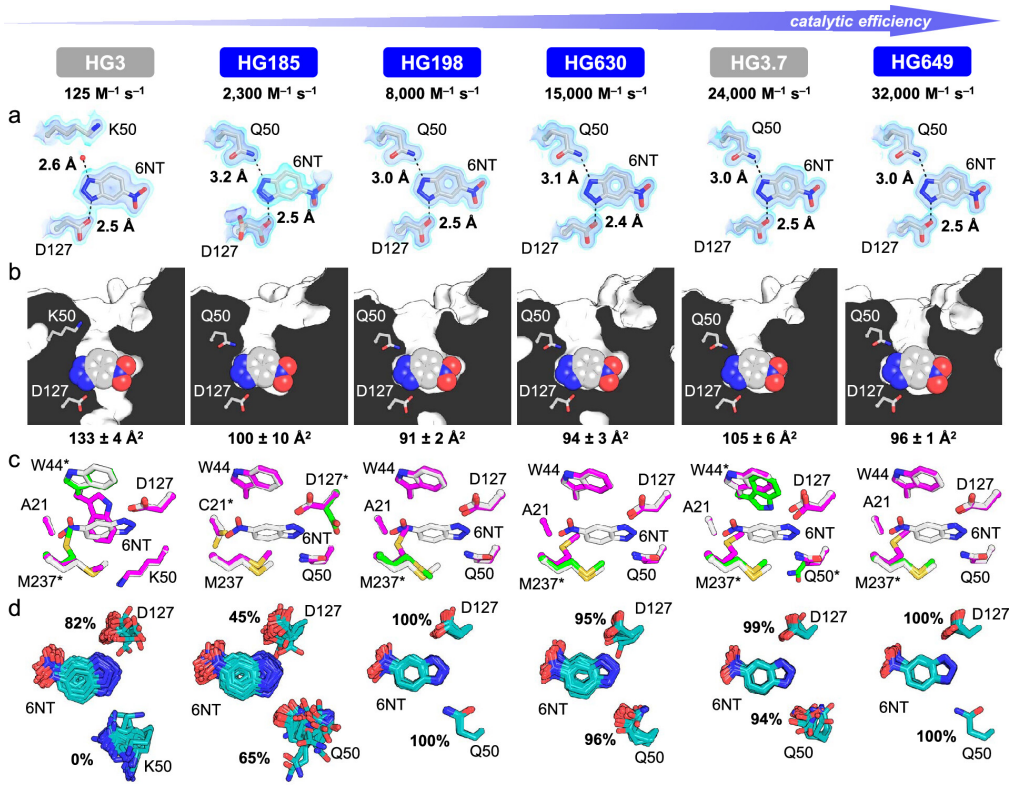


Figure 4.9: **Crystal structures of HG series Kemp eliminases.** In all cases, only atoms from chain A are shown. Structures of HG3 (PDB ID: 5RG4 and 5RGA) and its evolved variant HG3.7 (PDB ID: 5RG6 and 5RGC) were previously published [143]. **(a)** Hydrogen bonds to the 6-benzotriazole (6NT) TS analogue are shown as dashed lines, and distances are indicated. The red sphere represents a water molecule. The $2F_o - F_c$ map is shown in volume representation at two contour levels: 0.5 and 1.5 e Å^{-3} in light and dark blue, respectively. In the case of HG185, occupancy of 6NT in the crystal is approximately 60%, which is reflected by its lower electron density. **(b)** Cutaway view showing that active sites of designs are more complementary to the TS analogue than that of the HG3 template, as indicated by lower solvent-accessible surface areas for 6NT (mean ± s.d. across all chains). 6NT is shown as spheres. **(c)** Superposition of the 6NT-bound structure (white) with the highest (magenta) and lowest (green) occupancy conformers of the unbound structure for each Kemp eliminase. Active-site residues that are not preorganized are indicated by an asterisk. **(d)** Ensemble refinement of 6NT-bound structures shows that designed variants HG198, HG630, and HG649, as well as evolved variant HG3.7, display lower structural heterogeneity than HG3 and a higher percentage of ensemble members whose catalytic side chains (D127 and Q50) make hydrogen bonds with 6NT.

All designs bound 6NT in the same catalytically productive pose (Fig. 4.9a) as that observed in HG3, in excellent agreement with the computational models generated by ensemble-based design but in stark contrast with those generated using the average crystal structure (Fig. 4.10), highlighting the increased accuracy afforded by the use of a crystallographically derived ensemble [143]. In this pose, the acidic N–H bond of 6NT that mimics the cleavable C–H bond of the substrate is located within hydrogen-bonding distance to the carboxylate oxygen of D127, while the basic nitrogen atom corresponding to the phenolic oxygen of the TS forms a hydrogen bond with the Nε atom of Q50, similar to what was observed for evolved variant HG3.7 [143]. By contrast, HG3 does not form this second hydrogen bond because its K50 residue does not adopt a suitable side-chain conformation, and instead, it is a waternot molecule that hydrogen bonds to the basic nitrogen atom of 6NT. In all cases, angles defining the hydrogen bond between D127 and the TS analogue were improved or within error compared to HG3. In contrast to all other enzymes, the 6NT TS analogue of HG185 was refined to an occupancy of approximately 60%, indicating that its crystal comprises a mixture of enzyme molecules in their bound and unbound forms. Given that the soaking procedure we used to generate crystals of HG185 bound to 6NT was identical to the one used to soak isomorphous crystals of HG630 and HG649, which yielded 6NT bound at 100% occupancy, it is likely that the lower occupancy of this ligand in HG185 reflects its weaker binding in the active site, in agreement with the enzyme's 6.5–14-fold lower catalytic efficiency compared to HG630 and HG649 (Table 4.3). The lower occupancy of 6NT in the HG185 structure is reflected by the weaker electron density in the active site and the presence of an alternate D127 rotamer that does not hydrogen bond with the TS analogue (Fig. 4.9a). Regardless, the active site of all designed variants is more complementary to the TS than that of HG3, as indicated by the lower solvent-accessible surface area for 6NT when bound in the active site of these variants (Fig. 4.9b).

Next, we evaluated active-site preorganization, which is an important contributor to enzyme catalysis [151], by comparing structures in the presence and absence of bound 6NT. In HG3, the unbound form is not preorganized for catalysis as

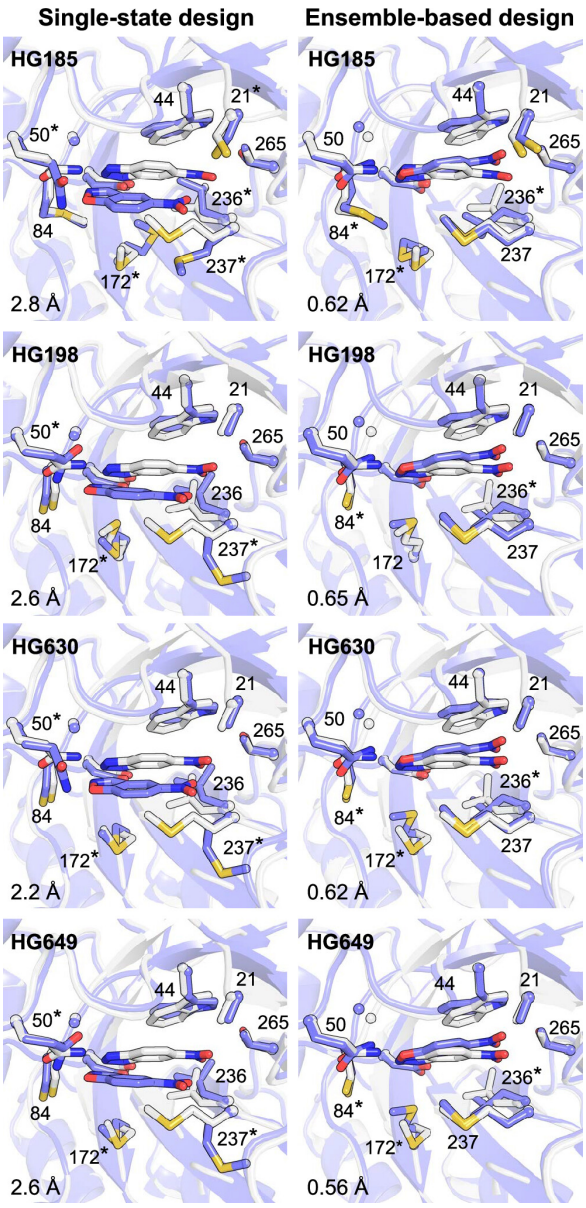


Figure 4.10: **Computational models vs crystal structures.** Crystal structures of HG-series Kemp eliminases with bound 6NT (white) are overlaid on the corresponding design models obtained by single-state design with the average crystal structure or ensemble-based design (blue). For clarity, only the major conformer of Chain A is shown. Root-mean-square deviations between the TS in the design models and the TS analogue in corresponding crystal structures are indicated at the bottom left. The TS and TS analogue are shown at the center of each image. Side chains of all residues forming the binding pocket are shown with the exception of P45, which was omitted for clarity. The sphere shows the α carbon of G83. Asterisks indicate residues that adopt side-chain rotamers varying by >20° around one or more side-chain dihedrals between the design model and crystal structure.

both W44 and M237 adopt conformations that would prevent productive binding of 6NT in the major or minor conformations, respectively (Fig. 4.9c). This is also the case for evolved variant HG3.7 [143]. HG185 is also poorly preorganized in the unbound form because C21 exclusively adopts a rotamer that would clash with 6NT. Furthermore, the D127 catalytic residue of HG185 adopts a low occupancy (7 and 14% for chains A and B, respectively), catalytically nonproductive conformation in the unbound form that cannot interact favorably with 6NT, likely contributing to its inferior activity compared with the other variants. By contrast, active sites of higher activity variants HG198, HG630, and HG649 are correctly preorganized for catalysis in a substantial portion of the molecules in the crystal, with only M237 adopting a nonproductive conformation at 47–64% occupancy (Fig. 4.9c). Overall, crystal structures confirmed that the activity of designed HG variants was increased by several of the mechanisms known to enhance enzyme catalysis and that we aimed to reproduce using our filtering criteria.

Enrichment of catalytically competent substates

The ability of an enzyme to frequently sample productive substates in its conformational ensemble is another important feature of efficient catalysis [103, 144, 82, 152]. To evaluate this property for our designs, we performed ensemble refinement of their 6NT-bound structures and evaluated the proportion of ensemble members that make the designed catalytic contacts with the TS analogue. We observed that in HG3, the hydrogen bond between the catalytic base and 6NT is formed in 82% of ensemble members and no interaction is observed between the ligand and K50 (Fig. 4.9d). In HG185, interactions between 6NT and catalytic residues D127 and Q50 are present in 45 and 65% of ensemble members, respectively, likely due to the lower occupancy of 6NT. In the case of the most active variants HG198, HG630, and HG649, catalytic contacts to both D127 and Q50 are presented in $\geq$95% of ensemble members, and their ensembles are less structurally diverse than those of HG3 or HG185 (Figs. 4.9d and 4.11a). The higher percentage of hydrogen bonds between 6NT and catalytic residues, and narrower conformational ensembles of high-activity HG variants whose structures contain

TS analogues at 100% occupancy, confirm a population of catalytically competent substates higher than in the ensemble of HG3.

Next, we quantified the extent to which ensemble-based design was able to shift the conformational ensemble observed for HG3 by analyzing the crystallographic ensemble models of the 6NT-bound forms of HG3 and the efficient HG variants that bound this TS analogue with 100% occupancy (HG198, HG630, and HG649). First, we calculated the pairwise backbone root-mean-square deviations (RMSDs) for all pairs of structures within a single ensemble model and plotted their distributions (Fig. 4.11a). Comparing the more active variants to HG3 demonstrated that all of these improved enzymes have a narrower (more self-consistent) ensemble, whose members have a lower median pairwise RMSD, which decreases as catalytic efficiency increases (e.g., 0.22, 0.17, 0.12, and 0.09 Å for HG3, HG198, HG630, and HG649, respectively). Next, we wanted to ask whether the ensembles of the more active variants are converging on the most highly populated conformations in the HG3 ensemble or around a different subset of conformations that might provide a better active-site structure. To assess this, we calculated the pairwise RMSD between the single conformer structure of HG3 bound to 6NT (PDB ID: 5RGA [143]) and the individual members of various ensemble models (HG198, HG630, HG649, and HG3). We observed that the single conformer structure of HG3 has a lower median RMSD relative to members of the HG3 ensemble model (0.17 Å) than it does relative to structures present in the ensemble models corresponding to higher activity variants (0.19–0.23 Å, Fig. 4.11b). Taken together, these RMSD calculations suggest that our design procedure narrows the ensembles of the more active variants around conformations that are substantially different from those that are most frequently populated by HG3.

KE70 designs

To further test our ensemble-based enzyme design method, we applied it to another Kemp eliminase that was designed on a different protein scaffold. This enzyme, KE70, contains a His/Asp dyad that acts as catalytic base and a Ser hydrogen-bond donor to stabilize productive substrate binding [133]. To allow di-

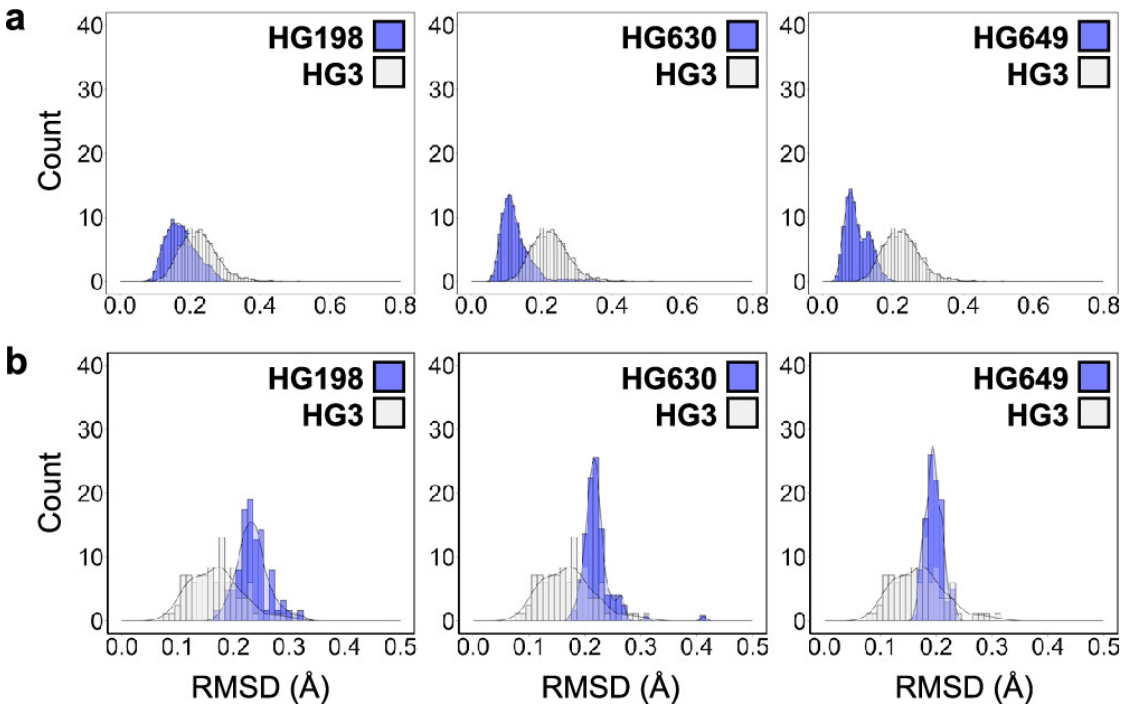


Figure 4.11: **Ensemble-based design leads to a shift in the conformational ensemble. (a)** Distributions of pairwise backbone RMSDs are calculated for all pairs of structures within a single ensemble model. Comparing the distributions derived from efficient HG variants (HG198, HG630, and HG649) bound to 6NT (blue) with those derived from the HG3 ensemble model, also with bound 6NT (white), reveals that the design process leads to a narrower (more self-consistent) distribution of conformations with a lower median pairwise RMSD. This effect is most pronounced for the most active variants (HG630 and HG649). **(b)** Distributions of pairwise RMSD between the single conformer structure of HG3 (PDB ID: 5RGA [143]) and the individual members of various ensemble models demonstrate that the ensembles of the improved variants have evolved to include conformations that are significantly different from HG3.

rect comparison to our HG enzymes, we designed variants using a theozyme comprising an Asp residue as a catalytic base and a Gln as a hydrogen-bond donor. Starting from the available unbound cryogenic crystal structure of KE70 (PDB ID: 3NPV [139]), we generated an 84-member ensemble and used it to design variants predicted to efficiently catalyze the Kemp elimination using a similar protocol as that used for HG enzymes but without the final filtering step involving MD (see Methods). Three variants were selected for kinetic analysis, and these contain between 11 and 13 mutations, including the D16 and Q138 catalytic motif as well as

two insertions in active-site loops (S20a and A240a) that we added after design. These loop insertions were manually introduced into KE70 during its directed evolution [139], and we added them to allow for direct comparison with evolved KE70 variants that contain these insertions, which increase the catalytic efficiency by 2.6-fold when introduced on their own into KE70 (Table 4.3). Two of the three tested designs were active. The most active variant, KE703, displays a k_{cat}/K_M value of 2500 ± 300 M^{-1} s^{-1}, which is 44-fold higher than KE70 and approximately 17-fold higher than that of a KE70 variant containing the two loop insertions (Table 4.3). To compare with evolved KE70 variants that contain the His/Asp dyad as a catalytic base, we introduced the D16H and N45D mutations into KE703, generating the KE703b variant. This variant has a k_{cat}/K_M value of 5500 ± 100 M^{-1} s^{-1}, which is approximately 100-fold higher than KE70 and 2-fold higher than that of the evolved variant R4 8/5A that was isolated after 4 rounds of directed evolution and screening of approximately 4000 clones [139] (Fig. 4.8c). KE703b shares a majority of its mutation sites with R4 8/5A, but its specific amino-acid substitutions are distinct, with only three mutations in common (Fig. 4.8d). These results confirm the ability of ensemble-based design to generate artificial enzymes that are orders of magnitude more active than the original *de novo* enzymes at a fraction of the experimental effort required by directed evolution. Furthermore, our results show that ensemble-based design does not require an ensemble generated from a room-temperature crystal structure with a bound TS analogue to be successful, which is advantageous given that such structures are less common than unbound ones solved at a cryogenic temperature.

4.2.3 Discussion

In this work, we used ensemble-based computational design to increase the catalytic efficiency of the HG3 and KE70 *de novo* Kemp eliminases by 2 orders of magnitude, yielding artificial enzymes with k_{cat}/K_M values that are only 4- or 10-fold lower than those of the most active variants evolved from these enzyme prototypes (e.g., k_{cat}/K_M of 126,000 and 53,100 M^{-1} s^{-1} for HG3.17 [143] and KE70 R8 12/12B [139], respectively). To achieve these results, we tested a number of

variants that are several orders of magnitude lower than those previously used to achieve comparable improvements by directed evolution [137, 139]. The success of our ensemble-based design relied on exploiting the spatiotemporally averaged information contained in X-ray crystallographic data to overcome the backbone sampling challenge that is responsible for much inaccuracy in enzyme design [137]. While others have previously attempted to address the backbone sampling issue by using ensembles generated via Monte Carlo-based backrub sampling and incorporating loop remodeling during the enzyme design process, activity improvements were modest (<5-fold) [139]. These results, in combination with our previous observation that use of an ensemble generated by unconstrained MD does not allow accurate recapitulation of catalytic contact and active-site configurations for an efficient artificial enzyme [143], suggest that to obtain large activity increases, it is not sufficient to introduce backbone flexibility in the design process. Instead, the correct backbone conformation must be sampled. Crystallographic ensemble refinement of the initial *de novo* enzyme satisfies this stringent requirement in that the resulting model represents the true conformational ensemble sampled by the enzyme, which should include productive substates that are better suited for stabilization of the TS than the average crystal structure. Generating the ensemble from a structure with a bound TS analogue likely helps to enrich such productive substates and could partially explain why the KE70 ensemble, which was derived from an unbound structure, led to a lower overall activity improvement. Although the use of an ensemble derived from enzymes in their bound and unbound forms led to activity improvements, one important disadvantage of our approach is that it requires a crystal structure of the original *de novo* enzyme to generate the ensemble. Since *de novo* enzymes are typically generated from small, monomeric protein scaffolds whose crystal structures have been determined to high resolution, we do not anticipate that structure determination will create a substantial bottleneck in most cases.

Our results support the idea that to design efficient artificial enzymes, it is essential to consider residual conformational heterogeneity present in the designs and whether it leads to the population of catalytically unproductive conformations. In

other words, dynamic aspects of enzyme structure, such as active-site preorganization and enrichment of catalytically competent substates, must also be considered. This is supported by the fact that the most active designs described here, HG630 and HG649, were ranked lower than less active ones such as HG44 and HG57 based on their predicted ability to stabilize the TS. Recently, researchers have proposed that enzyme design should be viewed as a population shift problem, where relative stabilities of conformational substates are redistributed by the introduction of mutations to stabilize productive substates [103, 153]. Previous work on the HG series of Kemp eliminases by our team [143] and others [103] demonstrated that directed evolution improves activity through essentially the same mechanism, with higher activity variants having reduced flexibility and greater active-site preorganization, both of which favor the population of catalytically active substates. The similarity is further highlighted by the fact that our ensemble-based design approach and directed evolution identify activity-enhancing mutations at many of the same positions.

Moving forward, we expect that ensemble-based design will find application in the creation of *de novo* enzymes for other organic transformations. However, for complex reactions involving multiple steps and transient species, additional developments will be required to make *de novo* enzyme design robust. For example, improvement of scoring functions used to evaluate sequences is necessary to accurately model forces that stabilize TSs in enzyme active sites, and implementation of multistate design approaches [45] for fine-tuning conformational transitions will be required to control progression through the catalytic cycle. We anticipate that ensemble-based design will be extended for achieving the latter objective as it could facilitate identification of sequences that stabilize multiple TSs by using multiple backbone templates from the ensemble. Additionally, we envision using crystallographic ensemble models to further enhance the population of catalytically competent conformations by identifying sequences that can stabilize productive substates while destabilizing nonproductive ones. The ability to remodel an enzyme conformational landscape [154] in this way, using negative design strategies to widen the energy gap between productive and nonproductive conformations, may

open the door to the design of more efficient artificial enzymes than previously possible. Beyond the scope of enzyme design, we expect that a similar strategy, which takes advantage of crystallographically derived ensemble models, could find utility in a broad range of protein design applications. For example, directed evolution following computational design of protein–protein interactions has been shown to enhance affinity by eliminating conformations from the ensemble that are nonproductive for binding [150]. Additionally, conformational heterogeneity is detrimental to the functions of computationally designed fluorescent proteins [155] and ligand-binding proteins [156]. Each of these observations suggests a potential application of our proposed approach.

4.2.4 Methods

Ensemble generation. Ensembles of backbone templates were generated using the ensemble refinement protocol [145] implemented in the PHENIX macromolecular structure determination software (v1.18.2–3874) [67]. In ensemble refinement, local atomic fluctuations are sampled by using MD simulations accelerated and restrained by electron density to produce ensemble models fitted to diffraction data. Briefly, input crystal structures were edited to remove low-occupancy conformers and assign an occupancy of 1.0 to the remaining conformer. Following the addition of riding hydrogens, parallel ensemble refinement simulations were performed using various combinations of the parameters p_{TLS} (0.6, 0.8, 0.9, 1.0), τ_x (0.5, 1.5, 2.0), and $w_{\text{X-ray}}$ (2.5, 5.0, 10.0), where p_{TLS} describes the percentage of atoms included in a translation-libration-screw (TLS) model used to remove the effects of global disorder, τ_x is the simulation time step, and $w_{\text{X-ray}}$ is the coupled temperature bath offset, which controls the extent to which the electron density contributes to the simulation force field such that the simulation runs at a target temperature of 300 K. The ensemble generated from each crystal structure that displayed the lowest R_{free} value was selected for design and analysis.

Computational enzyme design. All calculations were performed with the Triad protein design software [157] (Protabit, Pasadena, CA, USA) using a Monte Carlo with simulated annealing search algorithm for rotamer optimization. Indi-

vidual backbone templates from each structural ensemble were prepared for computational design using the `addH.py` application within Triad, which optimized and renamed hydrogen atoms, assigned protonation state to His residues, flipped His/Asn/Gln side chains if necessary, and rotated hydroxyl, sulfhydryl, methyl, and amino groups in Ser/Thr, Cys, Met, and Lys, respectively.

The Kemp elimination TS was built using parameters described by Privett and co-workers [113]. For HG3 designs, D127 was used as the catalytic base, and Gln/Asn/Thr/Ser was tested as hydrogen-bond donors at various positions. For KE70 designs, D16 served as the catalytic base (corresponding to position 17 in the KE70 crystal structure, PDB ID: 3NPV), Q138 as hydrogen-bond donor, and F48 was utilized for introduction of a pi-stacking interaction with the TS. For both HG3 and KE70 designs, all other positions in the active site were mutated to Gly during theozyme placement to avoid steric clashes with the designed catalytic residues. The 2002 Dunbrack backbone-independent rotamer library [115] with expansions of ± 1 standard deviation around χ_1 and χ_2 was used to provide side-chain conformations. A library of TS poses was generated in the active site by targeted ligand placement [131]. TS pose energies were calculated using a modified version of the Phoenix protein design energy function [157] consisting of a Lennard-Jones 12–6 van der Waals term from the Dreiding II force field [116] with atomic radii scaled by 0.9, a direction-dependent hydrogen bond term with a well depth of 8.0 kcal mol^{-1} and an equilibrium donor–acceptor distance of 2.8 Å [117], and an electrostatic energy term modeled using Coulomb's law with a distance-dependent dielectric of 10. During the energy calculation step, TS–side-chain interaction energies were biased to favor interactions that satisfy contact geometries by applying an energy benefit of -100 kcal mol^{-1} as described by Lassila et al. [131].

Next, active-site repacking calculations were performed on the theozyme-containing templates. In these calculations, the TS was translated ± 0.4 Å in each Cartesian coordinate in 0.2 Å increments and rotated 10° about all three axes (origin at TS geometric center) in 5° increments for a total combinatorial rotation/translation search space of 15,625 possible poses. The identities of catalytic residues were fixed and allowed to sample all rotamers of that amino acid type. Residues that were

converted to Gly in the theozyme placement step were allowed to sample rotamers of selected amino acids to form a mostly hydrophobic active site. Side-chain–TS interaction energies were biased to favor contacts satisfying geometries as done during the theozyme placement step. Rotamer optimization was carried out using the search algorithm, rotamer library, and energy function described above.

After active-site repacking, we calculated energy differences between each repacked structure and the corresponding structure in the absence of the TS where all designed residues (including catalytic residues) were converted to Gly. This energy subtraction was done to allow a comparison of energies for sequences across multiple backbone templates. We also measured the angles defining the hydrogen-bond contact between the catalytic base and the TS of the top-scoring sequence obtained from each backbone template. Sequences from the HG3 and KE70 backbone template that gave the lowest energy top-ranked sequence with angles between its catalytic base and the TS within 3 or 3.5% error of the optimal geometries calculated for acetamide dimers [149], respectively, were selected for library design and cleaning.

Computational library design. Library design was performed with the CLEARSS algorithm [90] using the energies obtained after active-site repacking on the single template selected as described earlier. For a specific library size configuration, which is the specific number of amino acids at each position in the protein, the highest probability set of amino acids at each position (based on the sum of their Boltzmann weights) were included in the library. A combinatorial library of approximately 1000 sequences derived from the selected HG3 or KE70 ensemble member was generated. To ensure that the lowest energy rotameric configuration was obtained for each library sequence, we used the `cleanSequences.py` app in Triad to "clean" their structures by optimizing rotamers on the input backbone template using the active-site repacking protocol described above but without modification of the catalytic residue rotameric configuration. To compare energies of "cleaned" sequences, we calculated the energy difference between each optimized structure and the corresponding structure in the absence of the TS where all designed residues (including catalytic residues) were converted to Gly. These energies

are reported throughout the figures and text.

Design filtering. To decrease the number of sequences to test experimentally, "cleaned" structures were filtered by using geometrical criteria. We kept sequences with (i) a TS solvent-accessible surface area lower than 100 or 102 Å^2 for HG3 and KE70 designs, respectively, and (ii) ≥70% of all designed active-site residues predicted to be preorganized, which was determined as those that adopted the same rotamer in the absence and presence of TS. Furthermore, sequences that contained ≥5 Met residues or ≥2 Cys residues were discarded to reduce conformational heterogeneity of the active site and prevent formation of disulfide bonds, respectively. A final filtering step consisting of short MD simulations was performed for HG3 designs, as described later.

Unconstrained MD simulation for filtering of HG3 designs. All simulations were performed using the Amber 2020 software (`http://ambermd.org/`) with the AMBER19SB force field [158]. Long-range electrostatics (>10 Å) were modeled using the particle mesh Ewald method [159], and a time step of 2 fs was used for the production phase. Parameters for the 5-nitrobenzisoxazole substrate were generated using the Antechamber [160] package. "Cleaned" computational models generated by Triad for the 55 sequences that passed the filtering procedure described earlier were prepared for MD. The TS structure was converted to that of the substrate, hydrogen atoms were added using Reduce [161], and each substrate-bound protein molecule was placed in a dodecahedral box with periodic boundary conditions where the distance between the protein surface and the box edges was set to 10 Å. After explicit TIP3P water molecules [162] were added, charges on protein atoms were neutralized with Na^+ and Cl^- counterions at a concentration of 0.15 M. The structures were then energy minimized with the steepest descent method to a target maximum force of 1000 kJ mol^{-1} nm^{-1}. Before equilibration, the system was heated to a target temperature of 300 K for 100 ps. The system was then equilibrated under an NVT ensemble for 100 ps at a temperature of 300 K using a Nosé-Hoover thermostat [163], while applying position restraints for heavy protein atoms. A second equilibration step under an NPT ensemble was performed for 100 ps with constant pressure and temperature of 1 bar and 300 K,

respectively, using the Berendsen barostat [164]. Following the removal of position restraints, 20 ns production runs under Parrinello–Rahman pressure coupling [165] were initiated from the final snapshot of the NPT equilibration. After MD, designs were evaluated and ranked according to their catalytic competency, which is the percentage of snapshots in which catalytic contacts between substrate and catalytic side chains were maintained.

Protein expression and purification. Codon-optimized (*E. coli*) and his-tagged (C-terminus) genes for Kemp eliminases cloned into the pET-29b(+) vector via NdeI and XhoI were obtained from Twist Bioscience. Enzymes were expressed in *E. coli* BL21-Gold (DE3) cells (Agilent) using lysogeny broth (LB) supplemented with 100 μg mL^{-1} kanamycin. Cultures were grown at 37 °C with shaking to an optical density at 600 nm of 0.3–0.7, and protein expression was initiated with 1 mM isopropyl-β-D-1-thiogalactopyranoside. Following incubation at 16 °C for 16 h with shaking (250 rpm), cells were harvested by centrifugation, resuspended in 8 mL of lysis buffer (5 mM imidazole in 100 mM potassium phosphate buffer pH 8.0) supplemented with 1.25 g mL^{-1} of lyophilized lysozyme and 1 U mL^{-1} of benzonase nuclease (Merck Millipore), and lysed with an EmulsiFlex-B15 cell disruptor (Avestin). Proteins were purified by immobilized metal affinity chromatography using Ni–NTA agarose (Qiagen) pre-equilibrated with lysis buffer in individual Econo-Pac gravity-flow columns (Bio-Rad). Columns were washed twice, first with 10 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0), and then with the same buffer containing 20 mM imidazole. Bound proteins were eluted with 250 mM imidazole in 100 mM potassium phosphate buffer (pH 8.0) and exchanged into 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride using Econo-Pac 10DG desalting prepacked gravity-flow columns (Bio-Rad). Proteins used for crystallization were further purified by gel filtration in 50 mM sodium citrate buffer (pH 5.5) and 150 mM sodium chloride using an ENrich SEC 650 size-exclusion chromatography column (Bio-Rad). Purified samples were concentrated using Amicon Ultracel-3K centrifugal filter units (EMD Millipore) and quantified by measuring the absorbance of the denatured protein in 6 M guanidinium chloride at 280 nm and applying Beer–Lambert's law

using calculated extinction coefficients obtained from the ExPAsy ProtParam tool.

Steady-state kinetics. All assays were carried out at 27 °C in 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride. Triplicate 200 μL reactions with varying concentrations of freshly prepared 5-nitrobenzisoxazole (Aablocks) dissolved in methanol (10% final concentration, pH of reaction mixture adjusted to 7.0 after addition of methanol-solubilized substrate) were initiated by the addition of 0.02–5.0 or 0.03–16.0 μM for HG or KE70 variants, respectively. Product formation was monitored spectrophotometrically at 380 nm (ε = 15,800 M^{-1} cm^{-1}) in individual 96-well plates (Greiner Bio-One) wells using a Biotek Synergy H1 multimode plate reader (Agilent). Path lengths for each well were calculated ratiometrically using the difference in absorbance of 100 mM sodium phosphate buffer (pH 7.0) supplemented with 100 mM sodium chloride and 10% methanol at 900 and 975 nm (27 °C). Linear phases of the kinetic traces were used to measure the initial reaction rates. For cases where saturation was not possible at the maximum substrate concentration tested (2 mM), data were fitted to the linear portion of the Michaelis–Menten model ($v_0 = (k_{cat}/K_M)[E_0][S]$), and k_{cat}/K_M was deduced from the slope. In all other cases, the data were fitted to the Michaelis–Menten equation to calculate individual k_{cat} and K_M parameters.

Crystallization. Enzyme variants were prepared in 50 mM sodium citrate buffer (pH 5.5). For samples that were cocrystallized with the TS analogue, a 100 mM stock solution of 6-nitrobenzotriazole (6NT, AstaTech) was prepared in dimethyl sulfoxide (DMSO) and diluted 20-fold in the enzyme solutions for a final concentration of 5 mM (5% DMSO). For each enzyme variant, we carried out initial crystallization trials in 15-well hanging-drop format using EasyXtal crystallization plates (Qiagen) and a crystallization screen that was designed to explore the chemical space around the crystallization conditions reported by Broom et al. [143]. Crystallization drops were prepared by mixing 1 μL of protein solution with 1 μL of the mother liquor and sealing the drop inside a reservoir containing an additional 500 μL of the mother liquor solution. This procedure was successful for HG185, HG198, and HG630; however, crystallization of HG649 required the identification of a new crystallization condition using sparse matrix screening.

We used commercial screening kits (NeXtal) to prepare crystallization drops in 96-well sitting drop plates by mixing 1 μL of the protein solution with 1 μL of the mother liquor and sealing the drop inside a reservoir containing an additional 100 μL of the mother liquor solution. After identifying a suitable mother liquor solution from screening experiments, large crystals suitable for X-ray diffraction experiments were obtained in a 15-well hanging-drop format as described for the other HG variants. HG185, HG630, and HG649 crystallized in the absence of the 6NT TS analogue, and the resulting crystals were isomorphous with the HG3 crystal form reported in the PDB (PDB ID: 5RGA and 5RG4). For these variants, crystals with bound 6NT were obtained by soaking the unbound crystals in mother liquor containing 10 mM 6NT for at least 30 min prior to data collection. On the other hand, HG198 could not be crystallized without 6NT present, and the crystals that were obtained belonged to space group $P1$ with three polypeptides in the asymmetric unit. To prepare crystals of HG198 without 6NT bound, we serially soaked 6NT-bound crystals in three drops of crystallization mother liquor without 6NT. Each sequential soaking step was carried out for at least 30 min. After the removal of 6NT by soaking, the $P1$ crystals undergo a transition to space group $P2_1$, with a smaller unit cell and only one molecule per asymmetric unit. Although we could not determine the cause of this unusual change in crystal symmetry upon soaking, we observed that it occurred with a high reproducibility.

X-ray data collection and processing. Prior to X-ray data collection, crystals were mounted in polyimide loops and sealed by using a MicroRT tubing kit (MiTeGen). Single-crystal X-ray diffraction data were collected on beamline 8.3.1 at the Advanced Light Source. The beamline was equipped with a Pilatus3 S 6M detector and was operated at a photon energy of 11,111 eV. Crystals were maintained at 280 K throughout the course of data collection. Each dataset was collected using a total X-ray dose of 50 kGy or less and covered a 180° wedge of reciprocal space. Multiple datasets were collected for each enzyme variant either from different crystals, or if their size permitted, from nonoverlapping volumes of larger crystals, in order to identify optimally diffracting samples. X-ray data were processed with the Xia2 0.5.492 program [166], which performed indexing, inte-

gration, scaling, and merging with DIALS [64, 167]. For each dataset, a resolution cutoff was taken where the $CC_{1/2}$ and $\langle I/\sigma I \rangle$ values for the merged intensities fell to approximately 0.5 and 1.0, respectively.

Structure determination. We obtained initial phase information for calculation of electron density maps by molecular replacement using the program Phaser [127], as implemented in v1.13.2998 of the PHENIX suite [128]. Several different models of HG variants, including crystal structures and design models, were used for molecular replacement. For all datasets except that of HG630 (+) 6NT, we performed iterative steps of manual model rebuilding followed by refinement of atomic positions, atomic displacement parameters, and occupancies using a TLS model, a riding hydrogen model, and automatic weight optimization. For HG630 (+) 6NT, refinement was done as described earlier but without the usage of a TLS model. All model building was performed using Coot 0.8.9.2 [168] and refinement steps were performed with `phenix.refine` within the PHENIX suite (v1.13–2998) [128, 66]. Restraints for 6NT were generated using `phenix.elbow` [129], starting from coordinates available in the Protein Data Bank (PDB ligand ID: 6NT).

Per-chapter acknowledgment

The text and figures of Part B of this chapter are adapted with permission from Rakotoharisoa, R. V.; Seifinoferest, B.; Zarifi, N.; Miller, J. D. M.; **Rodriguez, J. M.**; Thompson, M. C.; Chica, R. A. "Design of Efficient Artificial Enzymes Using Crystallographically Enhanced Conformational Sampling." *Journal of the American Chemical Society* **146**, 10001–10013 (2024). `https://doi.org/10.1021/jacs.4c00677`. Copyright 2024 American Chemical Society. The dissertation author was a co-author of this publication and was responsible for collecting room-temperature X-ray diffraction data on the designed Kemp eliminase variants.

Chapter 5

Conclusion

In this dissertation, I have addressed two major gaps in our understanding of interferon regulatory factor 3 (IRF3). First, I identified the first experimentally validated small-molecule binding sites on IRF3CTD through crystallographic fragment screening. These fragments provide starting points for structure-guided development of ligands targeting this previously undrugged transcription factor for both therapeutic and research applications. Second, I tested a mechanistically interesting but biophysically unvalidated proposal that the respiratory syncytial virus protein NS1 binds IRF3 directly, using orthogonal biophysical methods to conclude that specific binding cannot be detected under the tested conditions. This result underscores the need for biophysical validation of the many viral protein-IRF3 interactions reported on the basis of cellular evidence alone. In addition to the IRF3-focused work, I showed how crystallographic methods that reveal protein conformational heterogeneity can inform and improve computational protein design, yielding enzymes with significantly altered catalytic properties.

These projects sit within a broader shift in structural biology away from the single-structure paradigm that has historically defined macromolecular crystallography. For most of the field's history, the goal of a crystallographic experiment has been to produce a single, averaged atomic model from a cryo-cooled crystal. This representation captures the dominant conformation of a protein but suppresses the conformational and compositional heterogeneity that often determines function [169]. Increasingly, the field is reorienting toward statistical descriptions of struc-

ture: ensemble models that fit multiple conformers simultaneously to diffraction data [145, 170], multi-dataset event maps that recover partial-occupancy ligands invisible in any single dataset [38], and room-temperature and multi-temperature data collection that preserves dynamic information lost during cryo-cooling [169]. The projects in this dissertation are applications of this statistical turn—each treats heterogeneity not as noise to be averaged away, but as the information that addresses the biological or design problem at hand.

5.1 Crystallographic Fragment Screening Opens Doors for Ligand and Drug Development

Crystallographic fragment screening has emerged as a powerful approach to drug discovery over the past two decades, with methods and infrastructure now advanced to a level competitive with traditional high-throughput screening in both timeframe and throughput [171]. As a methodology, fragment screening is also a paradigm case of the statistical turn in crystallography: individual fragment hits are typically too weak and too partially occupied to be detected in any single dataset, and only emerge through pan-dataset density analysis across hundreds of essentially isomorphous crystals [38]. The method's success therefore depends fundamentally on treating crystallographic data as a statistical resource rather than as a route to a single representative structure. Despite this maturity, the method remains underutilized in many areas of drug discovery. The advantages that should drive broader adoption are substantial. Unlike the commonly used high-throughput screening, which is constrained to report on a functional read-out, fragment screening agnostically identifies multiple binding hotspots across the protein surface and generates information-rich datasets about ligand binding modes at each site. This becomes particularly valuable for targets like transcription factors that lack well-defined enzymatic activities, since the method does not require a functional assay [172]. Additionally, the structure-based nature of this method allows for the rational design of tightly-binding ligands through iterative optimization cycles. Beyond these practical advantages, fragment screening has

proven capable of yielding clinically effective compounds, with seven drugs developed from fragment hits having received FDA approval [171], the most recent being Capivasertib for breast cancer. The field should continue to evolve toward fuller exploitation of these capabilities, moving beyond the current landscape where only a fraction of druggable targets are explored through fragment screening.

IRF3 is a well-suited target for crystallographic fragment screening. As a transcription factor with no previously reported small-molecule ligands and no well-characterized druggable pockets, it represents the kind of undrugged target for which traditional structure-based drug discovery methods offer limited entry points. The identification of fragments and their binding modes provides concrete starting points for developing compounds that modulate IRF3 activation, serve as tools for mechanistic investigation, or probe host-pathogen interactions.

The results from our fragment screen enable multiple paths forward for developing IRF3 ligands. The site 1 fragments, which share a benzoic acid scaffold and engage the backbone carbonyl of W280 or Q279 through their carboxyl groups, are well-positioned for covalent elaboration toward C371 to create more potent ligands. Sites 2, 3, and 4 each contain multiple hits in distinct orientations that are amenable to fragment linking or merging strategies [173]. One fragment bound to a particularly important site, site 7, a common site of adaptor protein-binding and the target of rotavirus NSP1. Compounds targeting these sites may have therapeutic potential in cellular pathologies such as excessive inflammation, or as antivirals against pathogens like rotavirus. If any of these sites are a common target of viral antagonists, then compounds binding those sites may have broad-spectrum antiviral potential. Even before optimization to drug-like potency, elaborated compounds targeting these sites would provide valuable tools for solution-based characterization of IRF3 dimerization and its various protein-protein interactions. Chapter 3 demonstrated that a proposed interaction between RSV NS1 and IRF3, supported by cellular co-immunoprecipitation data, could not be validated as a specific binding interaction using purified components and orthogonal biophysical assays. This negative result does not rule out an interaction in the cellular context, but it highlights the necessity of biophysical validation before investing in

structural studies or further mechanistic characterization. The fragments identified here may aid in such validation, for example by serving as competitive probes in solution-binding assays for viral proteins proposed to engage the same surface. The fragment screening results thus provide not only direct chemical starting points for IRF3 ligand development but also potential tools for the broader characterization of IRF3 biology.

5.2 Computational Design Benefits from Crystallography

Historically, computational protein design has been formulated as a single-state problem, where the amino-acid sequence is optimized for a desired outcome within a fixed backbone template [174]. This single-state framing is poorly suited to problems whose outcome depends on the relative populations of multiple conformational states. Over the past decade, two complementary design strategies have been developed to address this limitation. Multistate design scores candidate sequences against multiple backbone templates simultaneously, allowing sequences to be optimized for preferential stabilization of one conformational state over another [45]. Ensemble-based design uses structural ensembles fitted to X-ray diffraction data as input templates, enabling the design procedure to identify and enrich catalytically productive conformations that are poorly represented in averaged structures [143]. Both strategies depend critically on accurate structural information about protein conformational heterogeneity, information that would be lost in routine structure determination.

The two collaborative projects presented in Chapter 4 demonstrate how information about conformational heterogeneity derived from crystallography can enable computational design to optimize for properties that have historically been ignored. The multi-state design of aspartate aminotransferase, informed by crystal structures of open and closed states, achieved targeted control over the conformational equilibrium of the enzyme, producing variants that preferentially favor either the closed or the open state. Similarly, ensemble-based design of the *de novo*

Kemp eliminases HG3 and KE70, using crystallographic ensembles as backbone inputs, produced enzymes with catalytic efficiencies comparable to those obtained through extensive directed evolution, without the need for high-throughput screening. Structural analysis confirmed that the activity gains arose from enrichment of catalytically competent substates, validating the design strategy. Together, these projects illustrate a key principle: when computational design incorporates accurate structural information about the dynamic nature of proteins, rather than treating them as static entities, it can shape conformational equilibria and substate populations in targeted ways.

These projects exemplify how specialized crystallographic methods that recover information about heterogeneity normally suppressed in routine structure determination can serve as a foundation for computational design. The identification and application of such information represents a broader methodological theme not only across the projects described in this dissertation, but within structural biology as a field. Crystallographic fragment screening identifies compositional heterogeneity, revealing multiple low-occupancy ligand binding sites that single-state structure determination would obscure. Multi-temperature crystallography and ensemble refinement reveal conformational heterogeneity, capturing the dynamic nature of protein structures that averaged models fail to represent. In both cases, the field's investment in statistical and multi-dataset methods has converted crystallography from a single-snapshot technique into a tool for characterizing the populations of conformers and ligand-bound states that govern protein function. Both applications demonstrate that strategic crystallographic approaches unlock structural information that downstream applications, whether drug discovery or computational design, require for success.

5.3 Future Applications

The work presented in this dissertation points toward several promising directions for future research. The fragment hits identified for IRF3CTD provide immediate starting points for developing lead-like drug molecules that modulate

IRF3 activation, whether to suppress excessive inflammation or to enhance antiviral immunity. Tool compounds may also be developed from these fragments and will enable systematic investigation of how ligand occupancy at each site affects IRF3 biology. The framework established for characterizing viral protein-IRF3 interactions through orthogonal biophysical methods can be applied to other viral antagonists, augmenting cellular evidence with quantitative binding measurements that provide a necessary foundation for mechanistic understanding and therapeutic development.

The success of computational design informed by crystallographic data suggests that the methodologies applied to AAT and the Kemp eliminases should find broader application. Many natural enzymes undergo conformational transitions during catalysis [175], and harnessing these transitions through design offers a path to improved tuning of function. The crystallographic methods that capture this information have been mature and accessible for well over a decade, but their integration into computational protein design workflows is comparatively recent. The Chica lab's approach—using ensemble outputs as input templates for design, then returning to crystallography to validate that the predicted conformational shifts occurred—exemplifies the kind of interdisciplinary loop that the field stands to gain from. As more design groups adopt this pattern, the field should expect continued improvements in the capability to design proteins for more efficient, more specific, and even new functions.

This dissertation demonstrates the power of using specialized crystallographic approaches to illuminate protein structure and dynamics in the service of biological and therapeutic problems. The maturation of methods for capturing compositional and conformational heterogeneity, combined with improving computational tools for integrating this information into drug discovery and protein design, promises to accelerate progress on historically difficult targets and to enable the rational design of proteins with properties optimized for specific applications. The work that follows should build on these foundations, applying the methodologies and insights discussed here to advance both fundamental understanding of protein biology and practical solutions to outstanding challenges in medicine and biotechnology.

Bibliography

[1] W C Au, P A Moore, W Lowther, Y T Juang, and P M Pitha. Identification of a member of the interferon regulatory factor family that binds to the interferon-stimulated response element and activates expression of interferon-induced genes. *Proceedings of the National Academy of Sciences*, 92(25):11657–11661, December 1995.

[2] Mariya AL Hamrashdi and Gareth Brady. Regulation of IRF3 activation in human antiviral signaling pathways. *Biochemical Pharmacology*, 200(February):115026, 2022.

[3] Thomas M Petro. IFN Regulatory Factor 3 in Health and Disease. *The Journal of Immunology*, 205(8):1981–1989, October 2020.

[4] Y.-T. Juang, W. Lowther, M. Kellum, W.-C. Au, R. Lin, J. Hiscott, and P. M. Pitha. Primary activation of interferon A and interferon B gene transcription by interferon regulatory factor 3. *Proceedings of the National Academy of Sciences*, 95(17):9837–9842, August 1998.

[5] Rongtuan Lin, Pierre Génin, Yaël Mamane, and John Hiscott. Selective DNA Binding and Association with the CREB Binding Protein Coactivator Contribute to Differential Activation of Alpha/Beta Interferon Genes by Interferon Regulatory Factors 3 and 7. *Molecular and Cellular Biology*, 20(17):6342–6353, 2000.

[6] Jason S. Rawlings, Kristin M. Rosler, and Douglas A. Harrison. The JAK/STAT signaling pathway. *Journal of Cell Science*, 117(8):1281–1283, March 2004.

[7] Saurabh Chattopadhyay and Ganes C. Sen. RIG-I-like receptor-induced IRF3 mediated pathway of apoptosis (RIPA): A new antiviral pathway. *Protein & Cell*, 8(3):165–168, March 2017.

[8] Sonam Popli, Sukanya Chakravarty, Shumin Fan, Anna Glanz, Siddhesh Aras, Laura E. Nagy, Ganes C. Sen, Ritu Chakravarti, and Saurabh Chattopadhyay. IRF3 inhibits nuclear translocation of NF-κB to prevent viral inflammation. *Proceedings of the National Academy of Sciences*, 119(37):e2121385119, September 2022.

[9] Rongtuan Lin, Christophe Heylbroeck, Paula M. Pitha, and John Hiscott. Virus-Dependent Phosphorylation of the IRF-3 Transcription Factor Regulates Nuclear Translocation, Transactivation Potential, and Proteasome-Mediated Degradation. *Molecular and Cellular Biology*, 18(5):2986–2996, May 1998.

[10] Bin Y. Qin, Cheng Liu, Suvana S. Lam, Hema Srinath, Rachel Delston, John J. Correia, Rik Derynck, and Kai Lin. Crystal structure of IRF-3 reveals mechanism of autoinhibition and virus-induced phosphoactivation. *Nature Structural Biology*, 10(11):913–921, 2003.

[11] Kiyohiro Takahasi, Nobuo N. Suzuki, Masataka Horiuchi, Mitsuaki Mori, Wakako Suhara, Yasutaka Okabe, Yukiko Fukuhara, Hiroaki Terasawa, Shizuo Akira, Takashi Fujita, and Fuyuhiko Inagaki. X-ray crystal structure of IRF-3 and its functional implications. *Nature Structural & Molecular Biology*, 10(11):922–927, November 2003.

[12] Tao Jing, Baoyu Zhao, Pengbiao Xu, Xinsheng Gao, Lei Chi, Huajun Han, Banumathi Sankaran, and Pingwei Li. The Structural Basis of IRF-3 Activation upon Phosphorylation. *The Journal of Immunology*, page ji2000026, 2020.

[13] Siqi Liu, Xin Cai, Jiaxi Wu, Qian Cong, Xiang Chen, Tuo Li, Fenghe Du, Junyao Ren, You Tong Wu, Nick V. Grishin, and Zhijian J. Chen. Phosphorylation of innate immune adaptor proteins MAVS, STING, and TRIF induces IRF3 activation. *Science*, 347(6227), 2015.

[14] Hans Häcker, Vanessa Redecke, Blagoy Blagoev, Irina Kratchmarova, Li-Chung Hsu, Gang G. Wang, Mark P. Kamps, Eyal Raz, Hermann Wagner, Georg Häcker, Matthias Mann, and Michael Karin. Specificity in Toll-like receptor signalling through distinct effector functions of TRAF3 and TRAF6. *Nature*, 439(7073):204–207, January 2006.

[15] Katherine A. Fitzgerald, Sarah M. McWhirter, Kerrie L. Faia, Daniel C. Rowe, Eicke Latz, Douglas T. Golenbock, Anthony J. Coyle, Sha-Mei Liao, and Tom Maniatis. IKKε and TBK1 are essential components of the IRF3 signaling pathway. *Nature Immunology*, 4(5):491–496, May 2003.

[16] Rashu B. Seth, Lijun Sun, Chee Kwee Ea, and Zhijian J. Chen. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-κB and IRF3. *Cell*, 122(5):669–682, 2005.

[17] Rashu B. Seth, Lijun Sun, and Zhijian J. Chen. Antiviral innate immunity pathways. *Cell Research*, 16(2):141–147, February 2006.

[18] Yasuo Tanaka and Zhijian J. Chen. STING Specifies IRF3 Phosphorylation by TBK1 in the Cytosolic DNA Signaling Pathway. *Science Signaling*, 5(214):ra20–ra20, March 2012.

[19] Mitsuaki Mori, Mitsutoshi Yoneyama, Takashi Ito, Kiyohiro Takahashi, Fuyuhiko Inagaki, and Takashi Fujita. Identification of Ser-386 of Interferon Regulatory Factor 3 as Critical Target for Inducible Phosphorylation That Determines Activation *. *Journal of Biological Chemistry*, 279(11):9698–9702, March 2004.

[20] Marc J. Servant, Nathalie Grandvaux, Benjamin R. TenOever, Delphine Duguay, Rongtuan Lin, and John Hiscott. Identification of the minimal phosphoacceptor site required for in vivo activation of interferon regulatory factor 3 in response to virus and double-stranded RNA. *Journal of Biological Chemistry*, 278(11):9441–9447, 2003.

[21] Bin Y. Qin, Cheng Liu, Hema Srinath, Suvana S. Lam, John J. Correia, Rik Derynck, and Kai Lin. Crystal structure of IRF-3 in complex with CBP. *Structure*, 13(9):1269–1277, 2005.

[22] Daniel Panne, Tom Maniatis, and Stephen C Harrison. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-β enhancer. *The EMBO Journal*, 23(22):4384–4393, November 2004.

[23] Baoyu Zhao, Chang Shu, Xinsheng Gao, Banumathi Sankaran, Fenglei Du, Catherine L. Shelton, Andrew B. Herr, Jun-Yuan Ji, and Pingwei Li. Structural basis for concerted recruitment and activation of IRF-3 by innate immune adaptor proteins. *Proceedings of the National Academy of Sciences*, 113(24):E3403–E3412, June 2016.

[24] I.P. Suarez, A. Cousido-Siah, A. Bonhoure, A. Mitschler, A. Podjarny, and G. Trave. Structure of HPV16 E6 oncoprotein in complex with IRF3 LxxLL motif: 6sja, September 2019.

[25] Line Lykke Andersen, Nanna Mørk, Line S. Reinert, Emil Kofod-Olsen, Ryo Narita, Sofie E. Jørgensen, Kristian A. Skipper, Klara Höning, Hans Henrik Gad, Lars Østergaard, Torben F. Ørntoft, Veit Hornung, Søren R. Paludan, Jacob Giehm Mikkelsen, Takashi Fujita, Mette Christiansen, Rune Hartmann, and Trine H. Mogensen. Functional IRF3 deficiency in a patient with herpes simplex encephalitis. *Journal of Experimental Medicine*, 212(9):1371–1379, July 2015.

[26] Kevin R. King, Aaron D. Aguirre, Yu-Xiang Ye, Yuan Sun, Jason D. Roh, Richard P. Ng, Rainer H. Kohler, Sean P. Arlauckas, Yoshiko Iwamoto, Andrej Savol, Ruslan I. Sadreyev, Mark Kelly, Timothy P. Fitzgibbons, Katherine A. Fitzgerald, Timothy Mitchison, Peter Libby, Matthias Nahrendorf,

and Ralph Weissleder. IRF3 and type I interferons fuel a fatal response to myocardial infarction. *Nature Medicine*, 23(12):1481–1487, December 2017.

[27] Manju Kumari, Xun Wang, Louise Lantier, Anna Lyubetskaya, Jun Eguchi, Sona Kang, Danielle Tenen, Hyun Cheol Roh, Xingxing Kong, Lawrence Kazak, Rasheed Ahmad, and Evan D. Rosen. IRF3 promotes adipose inflammation and insulin resistance and represses browning. *The Journal of Clinical Investigation*, 126(8):2839–2854, August 2016.

[28] Jingqian Su, Biyun Guan, Kunsen Chen, Zhihua Feng, Kai Guo, Xue Wang, Jianbin Xiao, Siyuan Chen, Wenzhi Chen, Long Chen, and Qi Chen. Fucoxanthin Attenuates Inflammation via Interferon Regulatory Factor 3 (IRF3) to Improve Sepsis. *Journal of Agricultural and Food Chemistry*, 71(33):12497–12510, August 2023.

[29] Shi Jiao, Jingmin Guan, Min Chen, Wenjia Wang, Chuanchuan Li, Yugong Wang, Yunfeng Cheng, and Zhaocai Zhou. Targeting IRF3 as a YAP agonist therapy against gastric cancer. *Journal of Experimental Medicine*, 215(2):699–718, 2018.

[30] Marie Bourdon, Caroline Manet, and Xavier Montagutelli. IRF3 in viral infections: More than just triggering the interferon response. *Genes & Immunity*, 26(5):419–428, October 2025.

[31] Cherrie D. Thompson, Bharati Matta, and Betsy J. Barnes. Therapeutic Targeting of IRFs: Pathway-Dependence or Structure-Based? *Frontiers in Immunology*, 9:2622, November 2018.

[32] Kirsten M. Spann, Kim C. Tran, and Peter L. Collins. Effects of Nonstructural Proteins NS1 and NS2 of Human Respiratory Syncytial Virus on Interferon Regulatory Factor 3, NF-κB, and Proinflammatory Cytokines. *Journal of Virology*, 79(9):5353–5362, May 2005.

[33] Junping Ren, Tianshuang Liu, Lan Pang, Kui Li, Roberto P. Garofalo, Antonella Casola, and Xiaoyong Bao. A novel mechanism for the inhibition of interferon regulatory factor-3-dependent gene expression by human respiratory syncytial virus NS1 protein. *Journal of General Virology*, 92(9):2153–2159, 2011.

[34] Peter Probst, John B. Grigg, Myra Wang, Ernesto Muñoz, Yueh-Ming Loo, Renee C. Ireton, Michael Gale, Shawn P. Iadonato, and Kristin M. Bedard. A small-molecule IRF3 agonist functions as an influenza vaccine adjuvant by modulating the antiviral immune response. *Vaccine*, 35(15):1964–1971, April 2017.

[35] Marie H. Foss, Susan L. Stevens, Haihong Jin, Elyse M. Allen, Dylan Nelson, Victor DeFilippis, Aaron Nilsen, and Mary P. Stenzel-Poore. Identification and Characterization of Small-Molecule IRF3-Dependent Immune Activators for Pharmaceutical Development. *ACS Chemical Biology*, 17(5):1073–1081, May 2022.

[36] Disha Patel, Joseph D. Bauman, and Eddy Arnold. Advantages of crystallographic fragment screening: Functional and mechanistic insights from a powerful platform for efficient drug discovery. *Progress in Biophysics and Molecular Biology*, 116(2):92–100, November 2014.

[37] Daren Fearon, Ailsa Powell, Alice Douangamath, Alexandre Dias, Charles W. E. Tomlinson, Blake H. Balcomb, Jasmin C. Aschenbrenner, Anthony Aimon, Isabel A. Barker, Felicity Bertram, José Brandão-Neto, Paul A. Coe, Patrick Collins, Louise E. Dunnett, Michael Fairhead, Richard J. Gildea, Mathew Golding, Tyler Gorrie-Stone, Paul V. Hathaway, Lizbé Koekemoer, Tobias Krojer, Ryan M. Lithgo, Elizabeth M. Maclean, Peter G. Marples, Halina Mikolajek, Xiaomin Ni, Kutumbarao H. V. Nidamarthi, Garry O'Donnell, Rachael Skyner, Romain Talon, Warren Thompson, Grant Watt, Conor F. Wild, Mark A. Williams, Max Winokan, Nathan David Wright, Graeme Winter, Elizabeth J. Shotton, and Frank von Delft. Accelerating Drug Discovery With High-Throughput Crystallographic Fragment Screening and Structural Enablement. *Applied Research*, 4(1):e202400192, 2025.

[38] Nicholas M. Pearce, Tobias Krojer, Anthony R. Bradley, Patrick Collins, Radosław P. Nowak, Romain Talon, Brian D. Marsden, Sebastian Kelm, Jiye Shi, Charlotte M. Deane, and Frank von Delft. A multi-crystal method for extracting obscured crystallographic states from conventionally uninterpretable electron density. *Nature Communications*, 8(1):15123, April 2017.

[39] Nicholas M. Pearce, Anthony R. Bradley, Tobias Krojer, Brian D. Marsden, Charlotte M. Deane, and Frank von Delft. Partial-occupancy binders identified by the Pan-Dataset Density Analysis method offer new chemical opportunities and reveal cryptic binding sites. *Structural Dynamics*, 4(3):032104, February 2017.

[40] Marion Schuller, Galen J. Correy, Stefan Gahbauer, Daren Fearon, Taiasean Wu, Roberto Efraín Díaz, Iris D. Young, Luan Carvalho Martins, Dominique H. Smith, Ursula Schulze-Gahmen, Tristan W. Owens, Ishan Deshpande, Gregory E. Merz, Aye C. Thwin, Justin T. Biel, Jessica K. Peters, Michelle Moritz, Nadia Herrera, Huong T. Kratochvil, QCRG Structural Biology Consortium, Anthony Aimon, James M. Bennett, Jose Brandao Neto, Aina E. Cohen, Alexandre Dias, Alice Douangamath, Louise Dunnett, Oleg Fedorov, Matteo P. Ferla, Martin R. Fuchs, Tyler J. Gorrie-Stone, James M.

Holton, Michael G. Johnson, Tobias Krojer, George Meigs, Ailsa J. Powell, Johannes Gregor Matthias Rack, Victor L. Rangel, Silvia Russi, Rachael E. Skyner, Clyde A. Smith, Alexei S. Soares, Jennifer L. Wierman, Kang Zhu, Peter O'Brien, Natalia Jura, Alan Ashworth, John J. Irwin, Michael C. Thompson, Jason E. Gestwicki, Frank von Delft, Brian K. Shoichet, James S. Fraser, and Ivan Ahel. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. *Science Advances*, 7(16):eabf8711, April 2021.

[41] Annette von Delft, Matthew D. Hall, Ann D. Kwong, Lisa A. Purcell, Kumar Singh Saikatendu, Uli Schmitz, John A. Tallarico, and Alpha A. Lee. Accelerating antiviral drug discovery: Lessons from COVID-19. *Nature Reviews Drug Discovery*, 22(7):585–603, July 2023.

[42] Tobias Krojer, James S. Fraser, and Frank von Delft. Discovery of allosteric binding sites by crystallographic fragment screening. *Current Opinion in Structural Biology*, 65:209–216, 2020.

[43] Andre Schutzer Godoy, Aline Minalli Nakamura, Alice Douangamath, Yun Song, Gabriela Dias Noske, Victor Oliveira Gawriljuk, Rafaela Sachetto Fernandes, Humberto D Muniz Pereira, Ketllyn Irene Zagato Oliveira, Daren Fearon, Alexandre Dias, Tobias Krojer, Michael Fairhead, Alisa Powell, Louise Dunnet, Jose Brandao-Neto, Rachael Skyner, Rod Chalk, Dávid Bajusz, Miklós Bege, Anikó Borbás, György Miklós Keserű, Frank von Delft, and Glaucius Oliva. Allosteric regulation and crystallographic fragment screening of SARS-CoV-2 NSP15 endoribonuclease. *Nucleic Acids Research*, 51(10):5255–5270, June 2023.

[44] Tatjana Barthel, Jan Wollenhaupt, Gustavo M.A. Lima, Markus C. Wahl, and Manfred S. Weiss. Large-Scale Crystallographic Fragment Screening Expedites Compound Optimization and Identifies Putative Protein-Protein Interaction Sites. *Journal of Medicinal Chemistry*, 2022.

[45] James A. Davey and Roberto A. Chica. Multistate approaches in computational protein design. *Protein Science*, 21(9):1241–1252, 2012.

[46] James J. Onuffer and Jack F. Kirsch. Redesign of the substrate specificity of escherichia coli aspartate aminotransferase to that of escherichia coli tyrosine aminotransferase by homology modeling and site-directed mutagenesis. *Protein Science*, 4(9):1750–1757, 1995.

[47] Vladimir N. Malashkevich, James J. Onuffer, Jack F. Kirsch, and Johan N. Jansonius. Alternating arginine-modulated substrate specificity in an engineered tyrosine aminotransferase. *Nature Structural Biology*, 2(7):548–553, July 1995.

[48] Louise Dalskov, Ryo Narita, Line L. Andersen, Nanna Jensen, Sonia Assil, Kennith H. Kristensen, Jacob G. Mikkelsen, Takashi Fujita, Trine H. Mogensen, Søren R. Paludan, and Rune Hartmann. Characterization of distinct molecular interactions responsible for IRF3 and IRF7 phosphorylation and subsequent dimerization. *Nucleic Acids Research*, 48(20):11421–11433, 2020.

[49] Tao Jing, Baoyu Zhao, Pengbiao Xu, Xinsheng Gao, Lei Chi, Huajun Han, Banumathi Sankaran, and Pingwei Li. The Structural Basis of IRF-3 Activation upon Phosphorylation. *The Journal of Immunology*, 205(7):1886–1896, October 2020.

[50] Raul S. Freitas, Tyler F. Crum, and Kislay Parvatiyar. SARS-CoV-2 Spike Antagonizes Innate Antiviral Immunity by Targeting Interferon Regulatory Factor 3. *Frontiers in Cellular and Infection Microbiology*, 11(January):1–8, 2022.

[51] Michael C. Thompson. Characterizing the interaction of HIV Vpu with human IRF-3 and establishment of human IRF-3 as a target for host-directed antiviral therapies. *HARC*, 1(December):1–6, 2020.

[52] Sang Yoon Park, Abdul A Waheed, Zai-Rong Zhang, Eric O Freed, and Juan S Bonifacino. HIV-1 Vpu accessory protein induces caspase-mediated cleavage of IRF3 transcription factor. *Journal of Biological Chemistry*, 289(51):35102–35110, 2014.

[53] Brian P. Doehle, Kristina Chang, Arjun Rustagi, John McNevin, M. Juliana McElrath, and Michael Gale. Vpu Mediates Depletion of Interferon Regulatory Factor 3 during HIV Infection by a Lysosome-Dependent Mechanism. *Journal of Virology*, 86(16):8367–8374, August 2012.

[54] Lucienne V. Ronco, Alla Y. Karpova, Marc Vidal, and Peter M. Howley. Human papillomavirus 16 E6 oncoprotein binds to interferon regulatory factor-3 and inhibits its transcriptional activity. *Genes & Development*, 12(13):2061–2072, January 1998.

[55] Santhana G. Devaraj, Nan Wang, Zhongbin Chen, Zihong Chen, Monica Tseng, Naina Barretto, Rongtuan Lin, Clarence J. Peters, Chien-Te K. Tseng, Susan C. Baker, and Kui Li. Regulation of IRF-3-dependent Innate Immunity by the Papain-like Protease Domain of the Severe Acute Respiratory Syndrome Coronavirus*. *Journal of Biological Chemistry*, 282(44):32208–32221, November 2007.

[56] Brian P. Doehle, Kristina Chang, Lamar Fleming, John McNevin, Florian Hladik, M. Juliana McElrath, and Michael Gale. Vpu-Deficient HIV Strains Stimulate Innate Immune Signaling Responses in Target Cells. *Journal of Virology*, 86(16):8499–8506, August 2012.

[57] Chenyang Yi, Zongzheng Zhao, Shengyu Wang, Xin Sun, Dan Zhang, Xiaomei Sun, Anding Zhang, and Meilin Jin. Influenza A Virus PA Antagonizes Interferon-β by Interacting with Interferon Regulatory Factor 3. *Frontiers in Immunology*, 8, September 2017.

[58] Hui Hui Wong, To Sing Fung, Shouguo Fang, Mei Huang, My Tra Le, and Ding Xiang Liu. Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. *Virology*, 515:165–175, February 2018.

[59] Hongjie Xia, Huanle Luo, Chao Shan, Antonio E. Muruato, Bruno T. D. Nunes, Daniele B. A. Medeiros, Jing Zou, Xuping Xie, Maria Isabel Giraldo, Pedro F. C. Vasconcelos, Scott C. Weaver, Tian Wang, Ricardo Rajsbaum, and Pei-Yong Shi. An evolutionary NS1 mutation enhances Zika virus evasion of host interferon induction. *Nature Communications*, 9(1):414, January 2018.

[60] Xinmeng Guan, Mudan Zhang, Ming Fu, Sukun Luo, and Qinxue Hu. Herpes Simplex Virus Type 2 Immediate Early Protein ICP27 Inhibits IFN-β Production in Mucosal Epithelial Cells by Antagonizing IRF3 Activation. *Frontiers in Immunology*, 10, February 2019.

[61] Raul S. Freitas, Tyler F. Crum, and Kislay Parvatiyar. SARS-CoV-2 Spike Antagonizes Innate Antiviral Immunity by Targeting Interferon Regulatory Factor 3. *Frontiers in Cellular and Infection Microbiology*, 11, January 2022.

[62] Fangjiao Huang, Xiaoli Han, Xiaohui Xiao, and Jinming Zhou. Covalent Warheads Targeting Cysteine Residue: The Promising Approach in Drug Development. *Molecules*, 27(22):7728, November 2022.

[63] Matteo P. Ferla, Rubén Sánchez-García, Rachael E. Skyner, Stefan Gahbauer, Jenny C. Taylor, Frank von Delft, Brian D. Marsden, and Charlotte M. Deane. Fragmenstein: Predicting protein–ligand structures of compounds derived from known crystallographic fragment hits using a strict conserved-binding–based methodology. *Journal of Cheminformatics*, 17(1):4, January 2025.

[64] G. Winter, D. G. Waterman, J. M. Parkhurst, A. S. Brewster, R. J. Gildea, M. Gerstel, L. Fuentes-Montero, M. Vollmar, T. Michels-Clark, I. D. Young, N. K. Sauter, and G. Evans. DIALS: Implementation and evaluation of a new integration package. *Acta Crystallographica Section D: Structural Biology*, 74(2):85–97, February 2018.

[65] G. Winter. Xia2: An expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography*, 43(1):186–190, February 2010.

[66] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, and P. D. Adams. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D: Biological Crystallography*, 68(4):352–367, April 2012.

[67] D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L.-W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, and P. D. Adams. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology*, 75(10):861–877, October 2019.

[68] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of *Coot*. *Acta Crystallographica Section D Biological Crystallography*, 66(4):486–501, April 2010.

[69] J. Agirre, M. Atanasova, H. Bagdonas, C. B. Ballard, A. Baslé, J. Beilsten-Edmands, R. J. Borges, D. G. Brown, J. J. Burgos-Mármol, J. M. Berrisford, P. S. Bond, I. Caballero, L. Catapano, G. Chojnowski, A. G. Cook, K. D. Cowtan, T. I. Croll, J. É Debreczeni, N. E. Devenish, E. J. Dodson, T. R. Drevon, P. Emsley, G. Evans, P. R. Evans, M. Fando, J. Foadi, L. Fuentes-Montero, E. F. Garman, M. Gerstel, R. J. Gildea, K. Hatti, M. L. Hekkelman, P. Heuser, S. W. Hoh, M. A. Hough, H. T. Jenkins, E. Jiménez, R. P. Joosten, R. M. Keegan, N. Keep, E. B. Krissinel, P. Kolenko, O. Kovalevskiy, V. S. Lamzin, D. M. Lawson, A. A. Lebedev, A. G. W. Leslie, B. Lohkamp, F. Long, M. Malý, A. J. McCoy, S. J. McNicholas, A. Medina, C. Millán, J. W. Murray, G. N. Murshudov, R. A. Nicholls, M. E. M. Noble, R. Oeffner, N. S. Pannu, J. M. Parkhurst, N. Pearce, J. Pereira, A. Perrakis, H. R. Powell, R. J. Read, D. J. Rigden, W. Rochira, M. Sammito, F. Sánchez Rodríguez, G. M. Sheldrick, K. L. Shelley, F. Simkovic, A. J. Simpkin, P. Skubak, E. Sobolev, R. A. Steiner, K. Stevenson, I. Tews, J. M. H. Thomas, A. Thorn, J. T. Valls, V. Uski, I. Usón, A. Vagin, S. Velankar, M. Vollmar, H. Walden, D. Waterman, K. S. Wilson, M. D. Winn, G. Winter, M. Wojdyr, and K. Yamashita. The CCP4 suite: Integrative software for macromolecular crystallography. *Acta Crystallographica Section D: Structural Biology*, 79(6):449–461, June 2023.

[70] Mariya AL Hamrashdi and Gareth Brady. Regulation of IRF3 activation in human antiviral signaling pathways. *Biochemical Pharmacology*, 200:115026, June 2022.

[71] Srirupa Chatterjee, Priya Luthra, Ekaterina Esaulova, Eugene Agapov, Benjamin C. Yen, Dominika M. Borek, Megan R. Edwards, Anuradha Mittal, David S. Jordan, Parameshwar Ramanan, Martin L. Moore, Rohit V. Pappu, Michael J. Holtzman, Maxim N. Artyomov, Christopher F. Basler, Gaya K. Amarasinghe, and Daisy W. Leung. Structural basis for human respiratory syncytial virus NS1-mediated modulation of host responses. *Nature Microbiology*, 2(9):17101, June 2017.

[72] Phillip Zhu, Philip R. Gafken, Ryan A. Mehl, and Richard B. Cooley. A highly versatile expression system for the production of multiply phosphorylated proteins. *ACS chemical biology*, 14(7):1564–1572, July 2019.

[73] Christopher S. Theile, Martin D. Witte, Annet E.M. Blom, Lenka Kundrat, Hidde L. Ploegh, and Carla P. Guimaraes. Site-specific N-terminal labeling of proteins using sortase-mediated reactions. *Nature protocols*, 8(9):1800–1807, September 2013.

[74] S. J. Benkovic, G. G. Hammes, and S. Hammes-Schiffer. Free-energy landscape of enzyme catalysis. *Biochemistry*, 47:3317–3321, 2008.

[75] S. J. Kerns, R. V. Agafonov, Y.-J. Cho, F. Pontiggia, R. Otten, D. V. Pachov, S. Kutter, L. A. Phung, P. N. Murphy, V. Thai, T. Alber, M. F. Hagan, and D. Kern. The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.*, 22:124–131, 2015.

[76] D. D. Boehr, D. McElheny, H. J. Dyson, and P. E. Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313:1638–1642, 2006.

[77] K. Henzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007.

[78] E. Campbell, M. Kaltenbach, G. J. Correy, P. D. Carr, B. T. Porebski, E. K. Livingstone, L. Afriat-Jurnou, A. M. Buckle, M. Weik, F. Hollfelder, N. Tokuriki, and C. J. Jackson. The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.*, 12:944–950, 2016.

[79] J. A. Kaczmarski, M. C. Mahawaththa, A. Feintuch, B. E. Clifton, L. A. Adams, D. Goldfarb, G. Otting, and C. J. Jackson. Altered conformational sampling along an evolutionary trajectory changes the catalytic activity of an enzyme. *Nat. Commun.*, 11:5945, 2020.

[80] M. A. Maria-Solano, J. Iglesias-Fernandez, and S. Osuna. Deciphering the allosterically driven conformational ensemble in tryptophan synthase evolution. *J. Am. Chem. Soc.*, 141:13049–13056, 2019.

[81] N. Tokuriki and D. S. Tawfik. Protein dynamism and evolvability. *Science*, 324:203–207, 2009.

[82] M. A. Maria-Solano, E. Serrano-Hervas, A. Romero-Rivera, J. Iglesias-Fernandez, and S. Osuna. Role of conformational dynamics in the evolution of novel enzyme function. *Chem. Commun. (Camb.)*, 54:6622–6634, 2018.

[83] B. E. Clifton and C. J. Jackson. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem. Biol.*, 23:236–245, 2016.

[84] A. Okamoto, T. Higuchi, K. Hirotsu, S. Kuramitsu, and H. Kagamiyama. X-ray crystallographic study of pyridoxal 5'-phosphate-type aspartate aminotransferases from Escherichia coli in open and closed form. *J. Biochem.*, 116:95–107, 1994.

[85] M. M. Islam, M. Goto, I. Miyahara, H. Ikushiro, K. Hirotsu, and H. Hayashi. Binding of C5-dicarboxylic substrate to aspartate aminotransferase: implications for the conformational change at the transaldimination step. *Biochemistry*, 44:8218–8229, 2005.

[86] I. Miyahara, K. Hirotsu, H. Hayashi, and H. Kagamiyama. X-ray crystallographic study of pyridoxamine 5'-phosphate-type aspartate aminotransferases from Escherichia coli in three forms. *J. Biochem.*, 116:1001–1012, 1994.

[87] S. Hayward and R. A. Lee. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J. Mol. Graph. Model.*, 21:181–183, 2002.

[88] J. A. Davey and R. A. Chica. Multistate computational protein design with backbone ensembles. In *Methods Mol. Biol.*, volume 1529, pages 161–179. 2017.

[89] J. A. Davey and R. A. Chica. Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Proteins*, 82:771–784, 2014.

[90] B. D. Allen, A. Nisthal, and S. L. Mayo. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc. Natl. Acad. Sci. USA*, 107:19838–19843, 2010.

[91] R. A. Chica, M. M. Moore, B. D. Allen, and S. L. Mayo. Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc. Natl. Acad. Sci. USA*, 107:20257–20262, 2010.

[92] J. M. Goldberg and J. F. Kirsch. The reaction catalyzed by Escherichia coli aspartate aminotransferase has multiple partially rate-determining steps, while that catalyzed by the Y225F mutant is dominated by ketimine hydrolysis. *Biochemistry*, 35:5280–5291, 1996.

[93] J. S. Fraser, M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern, and T. Alber. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, 462:669–673, 2009.

[94] D. A. Keedy, H. van den Bedem, D. A. Sivak, G. A. Petsko, D. Ringe, M. A. Wilson, and J. S. Fraser. Crystal cryocooling distorts conformational heterogeneity in a model Michaelis complex of DHFR. *Structure*, 22:899–910, 2014.

[95] J. Jäger, M. Moser, U. Sauder, and J. N. Jansonius. Crystal structures of Escherichia coli aspartate aminotransferase in two conformations: comparison of an unliganded open and two liganded closed forms. *J. Mol. Biol.*, 239:285–305, 1994.

[96] S. Rhee, M. M. Silva, C. C. Hyde, P. H. Rogers, C. M. Metzler, D. E. Metzler, and A. Arnone. Refinement and comparisons of the crystal structures of pig cytosolic aspartate aminotransferase and its complex with 2-methylaspartate. *J. Biol. Chem.*, 272:17293–17302, 1997.

[97] J. T. Hammill, S. Miyake-Stoner, J. L. Hazen, J. C. Jackson, and R. A. Mehl. Preparation of site-specifically labeled fluorinated proteins for ^{19}F-NMR structural characterization. *Nat. Protoc.*, 2:2601–2607, 2007.

[98] S. Kuramitsu, K. Hiromi, H. Hayashi, Y. Morino, and H. Kagamiyama. Presteady-state kinetics of Escherichia coli aspartate aminotransferase catalyzed reactions and thermodynamic aspects of its substrate specificity. *Biochemistry*, 29:5469–5476, 1990.

[99] G. Sharma and E. A. First. Thermodynamic analysis reveals a temperature-dependent change in the catalytic mechanism of Bacillus stearothermophilus tyrosyl-tRNA synthetase. *J. Biol. Chem.*, 284:4179–4190, 2009.

[100] A. R. Buller, P. van Roye, J. K. B. Cahn, R. A. Scheele, M. Herger, and F. H. Arnold. Directed evolution mimics allosteric activation by stepwise tuning of the conformational ensemble. *J. Am. Chem. Soc.*, 140:7256–7266, 2018.

[101] F. Rago, D. Saltzberg, K. N. Allen, and D. R. Tolan. Enzyme substrate specificity conferred by distinct conformational pathways. *J. Am. Chem. Soc.*, 137:13876–13886, 2015.

[102] P. M. Khade, A. Kumar, and R. L. Jernigan. Characterizing and predicting protein hinges for mechanistic insight. *J. Mol. Biol.*, 432:508–522, 2020.

[103] R. Otten, R. A. P. Padua, H. A. Bunzel, V. Nguyen, W. Pitsawong, M. Patterson, S. Sui, S. L. Perry, A. E. Cohen, D. Hilvert, and D. Kern. How directed evolution reshapes the energy landscape in an enzyme to boost catalysis. *Science*, 370:1442–1446, 2020.

[104] N. S. Hong, D. Petrović, R. Lee, G. Grynćoł, S. C. L. Kamerlin, N. Tokuriki, and C. J. Jackson. The evolution of multiple active site configurations in a designed enzyme. *Nat. Commun.*, 9:3900, 2018.

[105] B. E. Clifton, J. A. Kaczmarski, P. D. Carr, M. L. Gerth, N. Tokuriki, and C. J. Jackson. Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nat. Chem. Biol.*, 14:542–547, 2018.

[106] M. Kaltenbach, J. R. Burke, M. Dindo, A. Pabis, F. S. Munsberg, A. Rabin, S. C. L. Kamerlin, J. P. Noel, and D. S. Tawfik. Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat. Chem. Biol.*, 14:548–555, 2018.

[107] J. S. Marvin and H. W. Hellinga. Manipulation of ligand binding affinity by exploitation of conformational coupling. *Nat. Struct. Biol.*, 8:795–798, 2001.

[108] K. Y. Wei, D. Moschidi, M. J. Bick, S. Nerli, A. C. McShan, L. P. Carter, P.-S. Huang, D. A. Fletcher, N. G. Sgourakis, S. E. Boyken, and D. Baker. Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. *Proc. Natl. Acad. Sci. USA*, 117:7208–7215, 2020.

[109] Chemical Computing Group ULC. Molecular Operating Environment (MOE), 2022.02. 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, Canada, 2023.

[110] A. C. Eliot and J. F. Kirsch. Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. *Annu. Rev. Biochem.*, 73:383–415, 2004.

[111] J. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21:1049–1074, 2000.

[112] S. G. Nash. A survey of truncated-Newton methods. *J. Comput. Appl. Math.*, 124:45–59, 2000.

[113] H. K. Privett, G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk, and S. L. Mayo. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. USA*, 109:3790–3795, 2012.

[114] B. D. Allen and S. L. Mayo. An efficient algorithm for multistate protein design based on FASTER. *J. Comput. Chem.*, 31:904–916, 2010.

[115] Jr. Dunbrack, R. L. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.*, 12:431–440, 2002.

[116] S. L. Mayo, B. D. Olafson, and W. A. Goddard. Dreiding – a generic force-field for molecular simulations. *J. Phys. Chem.*, 94:8897–8909, 1990.

[117] B. I. Dahiyat and S. L. Mayo. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA*, 94:10172–10177, 1997.

[118] D. Shortle. Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci.*, 12:1298–1302, 2003.

[119] T. A. Addington, R. W. Mertz, J. B. Siegel, J. M. Thompson, A. J. Fisher, V. Filkov, N. M. Fleischman, A. A. Suen, C. Zhang, and M. D. Toney. Janus: prediction and ranking of mutations required for functional interconversion of enzymes. *J. Mol. Biol.*, 425:1378–1389, 2013.

[120] K. L. Heckman and L. R. Pease. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat. Protoc.*, 2:924–932, 2007.

[121] G. G. Mironov, V. Logiou, P. Bobrowska, and M. V. Berezovski. Bioanalysis for biocatalysis: multiplexed capillary electrophoresis-mass spectrometry assay for aminotransferase substrate discovery and specificity profiling. *J. Am. Chem. Soc.*, 135:13728–13736, 2013.

[122] O. Ernst and T. Zor. Linearization of the Bradford protein assay. *J. Vis. Exp.*, 38:1918, 2010.

[123] J. J. Porter, H. S. Jang, M. M. Haque, D. J. Stuehr, and R. A. Mehl. Tyrosine nitration on calmodulin enhances calcium-dependent association and activation of nitric-oxide synthase. *J. Biol. Chem.*, 295:2203–2211, 2020.

[124] M. Newville, T. Stensitzki, D. B. Allen, and A. Ingargiola. LMFIT: Non-linear least-square minimization and curve-fitting for Python, 2014.

[125] D. A. Nicholson, B. Jia, and D. J. Nesbitt. Measuring excess heat capacities of deoxyribonucleic acid (DNA) folding at the single-molecule level. *J. Phys. Chem. B*, 125:9719–9726, 2021.

[126] J. Beilsten-Edmands, G. Winter, R. Gildea, J. Parkhurst, D. Waterman, and G. Evans. Scaling diffraction data in the DIALS software package: algorithms and new approaches for multi-crystal scaling. *Acta Crystallogr. Sect. D*, 76:385–399, 2020.

[127] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read. Phaser crystallographic software. *J. Appl. Crystallogr.*, 40:658–674, 2007.

[128] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.*, 66:213–221, 2010.

[129] N. W. Moriarty, R. W. Grosse-Kunstleve, and P. D. Adams. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr. D Biol. Crystallogr.*, 65:1074–1080, 2009.

[130] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank. a computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, 80:319–324, 1977.

[131] J. K. Lassila, H. K. Privett, B. D. Allen, and S. L. Mayo. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA*, 103:16710–16715, 2006.

[132] A. Zanghellini, L. Jiang, A. M. Wollacott, G. Cheng, J. Meiler, E. A. Althoff, D. Rothlisberger, and D. Baker. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.*, 15:2785–2794, 2006.

[133] D. Rothlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453:190–195, 2008.

[134] L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, III, D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker. De novo computational design of retro-aldol enzymes. *Science*, 319:1387–1391, 2008.

[135] J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk,

F. E. Michael, and D. Baker. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels–Alder reaction. *Science*, 329:309–313, 2010.

[136] S. Bjelic, L. G. Nivón, N. Çelebi-Ölçüm, G. Kiss, C. F. Rosewall, H. M. Lovick, E. L. Ingalls, J. L. Gallaher, J. Seetharaman, S. Lew, G. T. Montelione, J. F. Hunt, F. E. Michael, K. N. Houk, and D. Baker. Computational design of enone-binding proteins with catalytic activity for the Morita–Baylis–Hillman reaction. *ACS Chem. Biol.*, 8:749–757, 2013.

[137] R. Blomberg, H. Kries, D. M. Pinkas, P. R. Mittl, M. G. Grutter, H. K. Privett, S. L. Mayo, and D. Hilvert. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature*, 503:418–421, 2013.

[138] A. H. Yeh, C. Norn, Y. Kipnis, D. Tischer, S. J. Pellock, D. Evans, P. Ma, G. R. Lee, J. Z. Zhang, I. Anishchenko, B. Coventry, L. Cao, J. Dauparas, S. Halabiya, M. DeWitt, L. Carter, K. N. Houk, and D. Baker. De novo design of luciferases using deep learning. *Nature*, 614:774–780, 2023.

[139] O. Khersonsky, D. Rothlisberger, A. M. Wollacott, P. Murphy, O. Dym, S. Albeck, G. Kiss, K. N. Houk, D. Baker, and D. S. Tawfik. Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. *J. Mol. Biol.*, 407:391–412, 2011.

[140] O. Khersonsky, G. Kiss, D. Rothlisberger, O. Dym, S. Albeck, K. N. Houk, D. Baker, and D. S. Tawfik. Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci. USA*, 109:10358–10363, 2012.

[141] L. Giger, S. Caner, R. Obexer, P. Kast, D. Baker, N. Ban, and D. Hilvert. Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.*, 9:494–498, 2013.

[142] R. Obexer, A. Godina, X. Garrabou, P. R. Mittl, D. Baker, A. D. Griffiths, and D. Hilvert. Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.*, 9:50–56, 2017.

[143] Aron Broom, Rojo V. Rakotoharisoa, Michael C. Thompson, Niayesh Zarifi, Erin Nguyen, Nurzhan Mukhametzhanov, Lin Liu, James S. Fraser, and Roberto A. Chica. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nature Communications*, 11(1):4808, September 2020.

[144] H. A. Bunzel, J. L. R. Anderson, D. Hilvert, V. L. Arcus, M. W. van der Kamp, and A. J. Mulholland. Evolution of dynamical networks enhances catalysis in a designer enzyme. *Nat. Chem.*, 13:1017–1022, 2021.

[145] B Tom Burnley, Pavel V Afonine, Paul D Adams, and Piet Gros. Modelling dynamics in protein crystal structures by ensemble refinement. *eLife*, 1:e00311, December 2012.

[146] D. A. Keedy, J. S. Fraser, and H. van den Bedem. Exposing hidden alternative backbone conformations in X-ray crystallography using qFit. *PLoS Comput. Biol.*, 11:e1004507, 2015.

[147] E. J. Levin, D. A. Kondrashov, G. E. Wesenberg, and G. N. Phillips, Jr. Ensemble refinement of protein crystal structures: Validation and application. *Structure*, 15:1040–1052, 2007.

[148] M. C. Thompson. Combining temperature perturbations with X-ray crystallography to study dynamic macromolecules: A thorough discussion of experimental methods. *Methods Enzymol.*, 688:255–305, 2023.

[149] A. V. Morozov, T. Kortemme, K. Tsemekhman, and D. Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. USA*, 101:6946–6951, 2004.

[150] J. T. Biel, M. C. Thompson, C. N. Cunningham, J. E. Corn, and J. S. Fraser. Flexibility and design: Conformational heterogeneity along the evolutionary trajectory of a redesigned ubiquitin. *Structure*, 25:739–749, 2017.

[151] G. Jindal and A. Warshel. Misunderstanding the preorganization concept can lead to confusions about the origin of enzyme catalysis. *Proteins*, 85:2157–2161, 2017.

[152] F. Yabukarski, T. Doukov, M. M. Pinney, J. T. Biel, J. S. Fraser, and D. Herschlag. Ensemble-function relationships to dissect mechanisms of enzyme catalysis. *Sci. Adv.*, 8:eabn7738, 2022.

[153] S. Osuna. The challenge of predicting distal active site mutations in computational enzyme design. *WIREs Comput. Mol. Sci.*, 11:e1502, 2021.

[154] Antony D. St-Jacques, Joshua M. Rodriguez, Matthew G. Eason, Scott M. Foster, Safwat T. Khan, Adam M. Damry, Natalie K. Goto, Michael C. Thompson, and Roberto A. Chica. Computational remodeling of an enzyme conformational landscape for altered substrate selectivity. *Nature Communications*, 14(1):6058, September 2023.

[155] S. Legault, D. P. Fraser-Halberg, R. L. McAnelly, M. G. Eason, M. C. Thompson, and R. A. Chica. Generation of bright monomeric red fluorescent proteins via computational design of enhanced chromophore packing. *Chem. Sci.*, 13:1408–1418, 2022.

[156] C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson, B. L. Stoddard, and D. Baker. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, 501:212–216, 2013.

[157] F. S. Lee, A. G. Anderson, and B. D. Olafson. Benchmarking TriadAb using targets from the second antibody modeling assessment. *Protein Eng. Des. Sel.*, 36:gzad013, 2023.

[158] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99sb protein force field. *Proteins*, 78:1950–1958, 2010.

[159] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[160] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25:247–260, 2006.

[161] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285:1735–1747, 1999.

[162] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[163] D. J. Evans and B. L. Holian. The Nosé–Hoover thermostat. *J. Chem. Phys.*, 83:4069–4074, 1985.

[164] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[165] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

[166] G. Winter, C. M. Lobley, and S. M. Prince. Decision making in xia2. *Acta Crystallogr. D Biol. Crystallogr.*, 69:1260–1273, 2013.

[167] P. Evans. Scaling and assessment of data quality. *Acta Crystallogr. D Biol. Crystallogr.*, 62:72–82, 2006.

[168] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):486–501, 2010.

[169] James S. Fraser, Henry van den Bedem, Avi J. Samelson, P. Therese Lang, James M. Holton, Nathaniel Echols, and Tom Alber. Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proceedings of the National Academy of Sciences*, 108(39):16247–16252, September 2011.

[170] Blake T. Riley, Stephanie A. Wankowicz, Saulo H. P. de Oliveira, Gydo C. P. van Zundert, Daniel W. Hogan, James S. Fraser, Daniel A. Keedy, and Henry van den Bedem. qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps. *Protein Science*, 30(1):270–285, 2021.

[171] Jessica Watt, Mathew P. Martin, Jane A. Endicott, and Martin.E.M. Noble. Different applications and differentiated libraries for crystallographic fragment screening. *Current Opinion in Structural Biology*, 90:102982, February 2025.

[172] Daniel A. Erlanson, Stephen W. Fesik, Roderick E. Hubbard, Wolfgang Jahnke, and Harren Jhoti. Twenty years on: The impact of fragments on drug discovery. *Nature Reviews Drug Discovery*, 15(9):605–619, September 2016.

[173] Alexandre Bancet, Claire Raingeval, Thierry Lomberget, Marc Le Borgne, Jean-François Guichou, and Isabelle Krimm. Fragment Linking Strategies for Structure-Based Drug Design. *Journal of Medicinal Chemistry*, 63(20):11420–11435, October 2020.

[174] Bassil I. Dahiyat and Stephen L. Mayo. De Novo Protein Design: Fully Automated Sequence Selection. *Science*, 278(5335):82–87, October 1997.

[175] Katherine A. Henzler-Wildman, Vu Thai, Ming Lei, Maria Ott, Magnus Wolf-Watz, Tim Fenn, Ed Pozharski, Mark A. Wilson, Gregory A. Petsko, Martin Karplus, Christian G. Hübner, and Dorothee Kern. Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844, December 2007.